

Encyclopedia of Computational Mechanics

Volume 1 Fundamentals

Editors

Erwin Stein

Institute of Structural and Computational Mechanics, University of Hannover, Hannover, Germany

René de Borst

Koiter Institute Delft, Delft University of Technology, The Netherlands

Thomas J. R. Hughes

Institute for Computational Engineering and Sciences, The University of Texas at Austin, TX, USA



WILEY

Copyright © 2004 John Wiley & Sons, Ltd,

The Atrium,
Southern Gate,
Chichester,
West Sussex,
PO19 8SQ, England

Telephone (+44) 1243 779777
Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London, W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or e-mailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons, Inc., 111 River Street,
Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street,
San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12,
D-69469 Weinheim, Germany

John Wiley & Sons Australia, Ltd, 33 Park Road,
Milton, Queensland, 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01,
Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 5553 Dundas Street West, Suite 400,
Etobicoke, Ontario, Canada M9B 6H5

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Encyclopedia of Computational Mechanics/Editors-in-chief, Erwin Stein, René de Borst and Thomas J.R. Hughes.

p. cm.
Includes bibliographical references and index.
ISBN 0-470-84699-2 (cloth)
1. Mechanics, Applied—Mathematical models. I. Stein, Erwin. II. Borst, René de.
III. Hughes, Thomas J. R.
TA350.E53 2004
620.1'001'5118—dc22

2004015104

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-84699-2

Typeset in 10/12 pt Times by Laserwords Private Limited, Chennai, India.

Printed and bound by Grafos SA, Barcelona, Spain.

This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

VOLUME 1: FUNDAMENTALS

Contributors to Volume 1

Preface

1 Fundamentals: Introduction and Survey

Erwin Stein

- 1 Motivation and Scope
- 2 Stages of Development and Features of Computational Mechanics
- 3 Survey of the Chapters of Volume 1
- 4 What We Do Expect

2 Finite Difference Methods

Owe Axelsson

- 1 Introduction
- 2 Two-point Boundary Value Problems
- 3 Finite Difference Methods for Elliptic Problems
- 4 Finite Difference Methods for Parabolic Problems
- 5 Finite Difference Methods for Hyperbolic Problems
- 6 Convection–Diffusion Problems
- 7 A Summary of Difference Schemes
- References
- Further Reading

3 Interpolation in *h*-version Finite Element Spaces

Thomas Apel

- 1 Introduction
- 2 Finite Elements
- 3 Definition of Interpolation Operators
- 4 The Deny–Lions Lemma
- 5 Local Error Estimates for the Nodal Interpolant
- 6 Local Error Estimates for Quasi-Interpolants
- 7 Example for a Global Interpolation Error Estimate
- 8 Related Chapters
- References

4 Finite Element Methods

Susanne C. Brenner/Carsten Carstensen

- 1 Introduction
- 2 Ritz–Galerkin Methods for Linear Elliptic Boundary Value Problems

ix

xi

1

1

5

7

7

12

18

28

36

50

52

53

55

73

73

74

- 3 Finite Element Spaces
- 4 A Priori Error Estimates for Finite Element Methods
- 5 A Posteriori Error Estimates and Analysis
- 6 Local Mesh Refinement
- 7 Other Aspects
- Acknowledgments
- References

5 The *p*-version of the Finite Element Method

Barna Szabó/Alexander Düster/Ernst Rank

- 1 Introduction
- 2 Implementation
- 3 Convergence Characteristics
- 4 Performance Characteristics
- 5 Applications to Nonlinear Problems
- 6 Outlook
- Acknowledgments
- Notes
- References
- Further Reading

6 Spectral Methods

Claudio Canuto/Alfio Quarteroni

- 1 Introduction
- 2 Fourier Methods
- 3 Algebraic Polynomial Expansion
- 4 Algebraic Expansions on Triangles
- 5 Stokes and Navier–Stokes Equations
- 6 Advection Equations and Conservation Laws
- 7 The Spectral Element Method
- 8 The Mortar Method
- References

7 Adaptive Wavelet Techniques in Numerical Simulation

Albert Cohen/Wolfgang Dahmen/Ronald DeVore

- 1 Introduction
- 2 Wavelets
- 3 Evolution Problems – Compression of Flow Fields
- 4 Boundary Integral Equations – Matrix Compression
- 5 A New Adaptive Paradigm

| | |
|---|------------|
| 6 Construction of Residual Approximations and Complexity Analysis | 187 |
| Acknowledgment | 195 |
| Notes | 195 |
| References | 195 |
| 8 Plates and Shells: Asymptotic Expansions and Hierarchic Models | 199 |
| <i>Monique Dauge/Erwan Faou/Zohar Yosibash</i> | |
| 1 Introduction | 199 |
| 2 Multiscale Expansions for Plates | 202 |
| 3 Hierarchical Models for Plates | 207 |
| 4 Multiscale Expansions and Limiting Models for Shells | 211 |
| 5 Hierarchical Models for Shells | 218 |
| 6 Finite Element Methods in Thin Domains | 219 |
| Acknowledgments | 229 |
| Notes | 229 |
| References | 229 |
| Further Reading | 232 |
| 9 Mixed Finite Element Methods | 237 |
| <i>Ferdinando Auricchio/Franco Brezzi/Carlo Lovadina</i> | |
| 1 Introduction | 237 |
| 2 Formulations | 238 |
| 3 Stability of Saddle-Points in Finite Dimensions | 246 |
| 4 Applications | 257 |
| 5 Techniques for Proving the <i>Inf-Sup</i> Condition | 269 |
| 6 Related Chapters | 276 |
| References | 276 |
| 10 Meshfree Methods | 279 |
| <i>Antonio Huerta/Ted Belytschko/Sonia Fernández-Méndez/Timon Rabczuk</i> | |
| 1 Introduction | 279 |
| 2 Approximation in Meshfree Methods | 280 |
| 3 Discretization of Partial Differential Equations | 291 |
| 4 Radial Basis Functions | 300 |
| 5 Discontinuities | 300 |
| 6 Blending Meshfree Methods and Finite Elements | 303 |
| References | 306 |
| 11 Discrete Element Methods | 311 |
| <i>Nenad Bilić</i> | |
| 1 Introduction | 311 |
| 2 Basic Discrete Element Framework and Regularization of Nonsmooth Contact Conditions | 314 |
| 3 Characterization of Interacting Bodies and Contact Detection | 317 |
| 4 Imposition of Contact Constraints and Boundary Conditions | 321 |
| 5 Modeling of Block Deformability | 324 |
| 6 Translato Continuum/Discontinuum, Fragmentation in Discrete Element Methods | 329 |
| 7 Time Integration – Temporal Discretization, Energy Balance, and Discrete Element Implementation | 331 |
| 8 Associated Frameworks and Developments | 333 |
| References | 335 |
| Further Reading | 337 |
| 12 Boundary Element Methods: Foundation and Error Analysis | 339 |
| <i>G. C. Hsiao/W. L. Wendland</i> | |
| 1 Introduction | 339 |
| 2 Boundary Integral Equations | 340 |
| 3 Variational Formulations | 347 |
| 4 The Galerkin-BEM | 358 |
| 5 The Role of Sobolev Index | 366 |
| 6 Concluding Remarks | 371 |
| Acknowledgments | 371 |
| References | 371 |
| Further Reading | 373 |
| 13 Coupling of Boundary Element Methods and Finite Element Methods | 375 |
| <i>Ernst P. Stephan</i> | |
| 1 Introduction | 375 |
| 2 Symmetric Coupling of Standard Finite Elements and Boundary Elements | 377 |
| 3 Fast Solvers for the hp-version of FE/BE Coupling | 389 |
| 4 Least Squares FE/BE Coupling Method | 394 |
| 5 FE/BE Coupling for Interface Problems with Signorini Contact | 396 |
| 6 Applications | 403 |
| 7 Concluding Remarks | 408 |
| References | 409 |
| 14 Arbitrary Lagrangian–Eulerian Methods | 413 |
| <i>J. Donea/A. Huerta/J.-Ph. Ponthot/A. Rodríguez-Ferran</i> | |
| 1 Introduction | 413 |
| 2 Descriptions of Motion | 415 |
| 3 The Fundamental ALE Equation | 417 |
| 4 ALE Form of Conservation Equations | 419 |
| 5 Mesh-update Procedures | 420 |
| 6 ALE Methods in Fluid Dynamics | 422 |
| 7 ALE Methods in Nonlinear Solid Mechanics | 426 |
| References | 433 |
| 15 Finite Volume Methods: Foundation and Analysis | 439 |
| <i>Timothy Barth/Mario Ohlberger</i> | |
| 1 Introduction: Scalar Nonlinear Conservation Laws | 439 |
| 2 Finite Volume (FV) Methods for Nonlinear Conservation Laws | 442 |
| 3 Higher-order Accurate FV Generalizations | 450 |
| 4 Further Advanced Topics | 464 |
| 5 Concluding Remarks | 470 |
| 6 Related Chapters | 470 |
| References | 470 |
| 16 Geometric Modeling of Complex Shapes and Engineering Artifacts | 475 |
| <i>F.-E. Wolter/N. Peinecke/M. Reuter</i> | |
| 1 Architecture of Modeling Systems | 475 |
| 2 Voxel Representation | 476 |
| 3 Surface Patches | 477 |
| 4 Boundary Representation | 481 |
| 5 Constructive Solid Geometry | 483 |

| | |
|---|------------|
| 6 Media Modeling | 485 |
| 7 Attributes | 490 |
| 8 Outlook and Concluding Remarks | 492 |
| Acknowledgments | 494 |
| Notes | 494 |
| References | 494 |
| 17 Mesh Generation and Mesh Adaptivity | 497 |
| <i>P. L. George/H. Borouchaki/P. J. Frey/P. Laug/E. Saltel</i> | |
| 1 Introduction | 497 |
| 2 A Brief History | 498 |
| 3 Mesh-Generation Methods | 499 |
| 4 Quality Meshing and Adaptivity | 502 |
| 5 Adaptive FEM Computations | 510 |
| 6 Large-size Problem, Parallelism and Adaptivity | 516 |
| 7 Meshing for Moving Boundary Problems | 517 |
| 8 Application Examples | 519 |
| 9 Conclusions | 520 |
| References | 521 |
| Further Reading | 523 |
| 18 Computational Visualization | 525 |
| <i>William J. Schroeder/Mark S. Shephard</i> | |
| 1 Introduction | 525 |
| 2 Data Forms | 528 |
| 3 Visualization Algorithms | 531 |
| 4 Volume Rendering | 541 |
| 5 Methods in Large Data Visualization | 542 |
| 6 Taxonomy for Data Visualization Systems | 543 |
| 7 Interfacing the Computational System with the Visualization System | 546 |
| References | 548 |
| 19 Linear Algebraic Solvers and Eigenvalue Analysis | 551 |
| <i>Henk A. van der Vorst</i> | |
| 1 Introduction | 551 |
| 2 Mathematical Preliminaries | 553 |
| 3 Direct Methods for Linear Systems | 553 |
| 4 Preconditioning | 560 |
| 5 Incomplete LU Factorizations | 562 |
| 6 Methods for the Complete Eigenproblem | 567 |
| 7 Iterative Methods for the Eigenproblem | 571 |
| Notes | 574 |
| References | 575 |
| 20 Multigrid Methods for FEM and BEM Applications | 577 |
| <i>Wolfgang Hackbusch</i> | |
| 1 General Remarks on Multigrid Methods | 577 |
| 2 Two-Grid Iteration | 581 |
| 3 Multigrid Method | 584 |
| 4 Application to Finite Element Equations | 586 |
| 5 Additive Variant | 589 |
| 6 Nested Iteration | 590 |
| 7 Nonlinear Equations | 592 |
| 8 Eigenvalue Problems | 593 |
| 9 Applications to the Boundary Element Method (BEM) | 593 |
| References | 595 |
| Further Reading | 595 |
| 21 Panel Clustering Techniques and Hierarchical Matrices for BEM and FEM | 597 |
| <i>Wolfgang Hackbusch</i> | |
| 1 Introduction | 597 |
| 2 The Panel Clustering Method (First Version) | 600 |
| 3 The Panel Clustering Method (Second Version) | 606 |
| 4 Hierarchical Matrices | 607 |
| References | 615 |
| 22 Domain Decomposition Methods and Preconditioning | 617 |
| <i>V. G. Korneev/U. Langer</i> | |
| 1 Introduction | 617 |
| 2 Domain Decomposition History | 619 |
| 3 Fundamentals of Schwarz's Methods | 621 |
| 4 Overlapping Domain Decomposition Methods | 630 |
| 5 Nonoverlapping Domain Decomposition Methods | 633 |
| Acknowledgments | 644 |
| References | 644 |
| Further Reading | 647 |
| 23 Nonlinear Systems and Bifurcations | 649 |
| <i>Werner C. Rheinboldt</i> | |
| 1 Introduction | 649 |
| 2 General Iterative Processes | 650 |
| 3 Some Classes of Iterative Methods | 657 |
| 4 Parameterized Systems | 661 |
| Bifurcation | 669 |
| References | 673 |
| 24 Adaptive Computational Methods for Parabolic Problems | 675 |
| <i>K. Eriksson/C. Johnson/A. Logg</i> | |
| 1 What is a Parabolic Problem? | 675 |
| 2 Outline | 676 |
| 3 References to the Literature | 676 |
| 4 Introduction to Adaptive Methods for IVPs | 677 |
| 5 Examples of Stiff IVPs | 680 |
| 6 A Nonstiff IVP: The Lorenz System | 683 |
| 7 Explicit Time-stepping for Stiff IVPs | 683 |
| 8 Strong Stability Estimates for an Abstract Parabolic Model Problem | 686 |
| 9 Adaptive Space–Time Galerkin Methods for the Heat Equation | 689 |
| 10 A Priori and A Posteriori Error Estimates for the Heat Equation | 690 |
| 11 Adaptive Methods/Algorithms | 691 |
| 12 Reliability and Efficiency | 691 |
| 13 Strong Stability Estimates for the Heat Equation | 691 |
| 14 A Priori Error Estimates for the L_2 - and Elliptic Projections | 692 |
| 15 Proof of the A Priori Error Estimates | 693 |
| 16 Proof of the A Posteriori Error Estimates | 695 |
| 17 Extension to Systems of Convection–Diffusion–reaction Problems | 696 |

| | | | |
|--|-----|---|-----|
| 18 Examples of Reaction-Diffusion Problems | 696 | 26 Finite Element Methods for Maxwell Equations | 723 |
| 19 Comparison with the Standard Approach to Time Step Control for ODEs | 699 | <i>Leszek Demkowicz</i> | |
| 20 Software | 702 | 1 Maxwell Equations | 723 |
| References | 702 | 2 Variational Formulation | 725 |
| Further Reading | 702 | 3 Exact Sequences | 727 |
| | | 4 Projection-based Interpolation. De Rham Diagram | 732 |
| | | 5 Additional Comments | 734 |
| | | 6 Related Chapters | 735 |
| | | Acknowledgment | 735 |
| | | Notes | 735 |
| | | References | 736 |
| | | Further Reading | 737 |
| 25 Time-dependent Problems with the Boundary Integral Equation Method | 703 | | |
| <i>Martin Costabel</i> | | | |
| 1 Introduction | 703 | | |
| 2 Space-time Integral Equations | 705 | | |
| 3 Laplace Transform Methods | 713 | | |
| 4 Time-stepping Methods | 714 | | |
| References | 719 | | |
| | | Contents for Volumes 2 and 3 | 739 |
| | | Subject Index | 745 |

Contributors to Volume 1

Thomas Apel
Universität der Bundeswehr München, Neubiberg, Germany

Ferdinando Auricchio
Università di Pavia and IMATI-C.N.R., Pavia, Italy

Owe Axelsson
University of Nijmegen, Nijmegen, The Netherlands

Timothy Barth
NASA Ames Research Center, Moffett Field, CA, USA

Ted Belytschko
Northwestern University, Evanston, IL, USA

Nenad Bicanic
University of Glasgow, Glasgow, Scotland

H. Borouchaki
Université de Technologie de Troyes, Troyes Cedex, France

Susanne C. Brenner
University of South Carolina, Columbia, SC, USA

Franco Brezzi
Università di Pavia and IMATI-C.N.R., Pavia, Italy

Claudio Canuto
Politecnico di Torino, Turin, Italy

Carsten Carstensen
Humboldt-Universität zu Berlin, Berlin, Germany

Albert Cohen
Université Pierre et Marie Curie, Paris, France

Martin Costabel
IRMAR, Université de Rennes 1, Campus de Beaulieu, Rennes, France

Wolfgang Dahmen
Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Germany

Monique Dauge
IRMAR, Université de Rennes 1, Campus de Beaulieu, Rennes, France

Leszek Demkowicz
The University of Texas at Austin, Austin, TX, USA

Ronald DeVore
University of South Carolina, Columbia, SC, USA

Jean Donea[†]
Université de Liège, Liège, Belgium

Alexander Düster
Lehrstuhl für Bauinformatik, Technische Universität München, Munich, Germany

K. Eriksson
Chalmers University of Technology, Göteborg, Sweden

Erwan Faou
INRIA Rennes, Campus de Beaulieu, Rennes, France

Sonia Fernández-Méndez
Universitat Politècnica de Catalunya, Barcelona, Spain

P. J. Frey
INRIA, Projet Gamma, Domaine de Voluceau, Rocquencourt, Le Chesnay Cedex, France

[†] Jean Donea sadly passed away in June 2004.

P. L. George
INRIA, *Projet Gamma, Domaine de Voluceau, Rocquencourt, Le Chesnay Cedex, France*

Wolfgang Hackbusch
Max-Planck-Institut für Mathematik in den Naturwissenschaften, Inselstr., Leipzig, Germany

G. C. Hsiao
University of Delaware, Newark, DE, USA

Antonio Huerta
Universitat Politècnica de Catalunya, Barcelona, Spain

C. Johnson
Chalmers University of Technology, Göteborg, Sweden

V. G. Koruev
St. Petersburg State Polytechnical University, St. Petersburg, Russia, and University of Westminster, London, UK

U. Langer
Johannes Kepler University Linz, Linz, Austria

P. Laug
INRIA, *Projet Gamma, Domaine de Voluceau, Rocquencourt, Le Chesnay Cedex, France*

A. Logg
Chalmers University of Technology, Göteborg, Sweden

Carlo Lovadina
Università di Pavia and IMATI-C.N.R., Pavia, Italy

Mario Ohlberger
Freiburg University, Freiburg, Germany and University of Maryland, College Park, MD, USA

N. Peinecke
University of Hannover, Hannover, Germany

J.-Ph. Ponthot
Université de Liège, Liège, Belgium

Alfio Quarteroni
MOX, Politecnico di Milano, Milan, Italy and IACS, School of Mathematics, EPFL, Lausanne, Switzerland

Timon Rabczuk
Northwestern University, Evanston, USA

Ernst Rank
Lehrstuhl für Bauinformatik, Technische Universität München, Munich, Germany

M. Reuter
University of Hannover, Hannover, Germany

Werner C. Rheinboldt
University of Pittsburgh, Pittsburgh, PA, USA

A. Rodríguez-Ferran
Universitat Politècnica de Catalunya, Barcelona, Spain

E. Salte
INRIA, *Projet Gamma, Domaine de Voluceau, Rocquencourt, Le Chesnay Cedex, France*

William J. Schroeder
Kitware, Inc., Clifton Park, NY, USA

Mark S. Shephard
Rensselaer Polytechnic Institute, Troy, NY, USA

Erwin Stein
University of Hannover, Hannover, Germany

Ernst P. Stephan
Institut für Angewandte Mathematik, Universität Hannover, Hannover, Germany

Barna Szabó
Washington University, St. Louis, MO, USA

Henk A. van der Vorst
Utrecht University, Utrecht, The Netherlands

W. L. Wendland
Universität Stuttgart, Stuttgart, Germany

F.-E. Wolter
University of Hannover, Hannover, Germany

Zohar Yosibash
Department of Mechanical Engineering, Ben-Gurion University, Beer Sheva, Israel

Preface

After nearly half a century of developments in numerical methods, the field of computational mechanics has become sufficiently mature to collect the achievements and summarize the state-of-the-art in a comprehensive, authoritative major reference work. This idea, first conceived in 1999, has resulted in the *Encyclopedia of Computational Mechanics*. It has been the intention of the editors and the publisher to provide the community with a systematic, well-organized survey of established as well as recently developed computational methods, covering applied and computational mathematics, computer science, the various branches of solid and fluid mechanics, and all the available discretization methods. Attention has also been paid to many engineering and other applications.

We have invited first-class scientists and engineers to join us in this challenging endeavor. It is our pleasure to thank all our contributors warmly for their excellent chapters and for completing them in a tight time frame.

A major reference work like the *Encyclopedia of Computational Mechanics* should provide trustworthy facts and information. We hope that the reviewing process to which each chapter has been submitted will guarantee the high quality we have aimed for. We would like to acknowledge the careful and constructive work of all of our reviewers.

The *Encyclopedia of Computational Mechanics* is organized in three volumes with the subjects 'Fundamentals' (Volume 1), 'Solids and Structures' (Volume 2), and 'Fluids' (Volume 3), and is published both in print and on line.

Volume 1, Fundamentals, contains contributions related to mathematics, mechanics, and computer science, and is structured as discretization methods (fourteen chapters), treating approximations with finite differences, discrete variational forms, boundary integral equations and further problem-oriented techniques, and the generation and visualization of geometry; meshes and results (three chapters); various direct and iterative solvers (five chapters); and time-dependent problems (three chapters).

Volume 2, Solids and Structures, is organized into five different parts, namely, structural behavior (four chapters);

constitutive theories and their discretization via finite element or boundary element methods (seven chapters); materials and processing (five chapters); interaction problems (five chapters); and identification, stochastic, and optimization (two chapters).

Volume 3, Fluids, builds on the fundamentals described in Volume 1. The chapters in Volume 3 fall within four main groupings. The first (four chapters) includes chapters describing additional basic methodologies used in computational fluid dynamics. The second (seven chapters) comprises chapters on various aspects of incompressible viscous flows. The third (four chapters) focuses on compressible fluid dynamics. The fourth (two chapters) pertains to problems involving moving domains and free surfaces.

Returning to the question 'Why the *Encyclopedia of Computational Mechanics* now?', we believe that the field of computational mechanics has now reached a high degree of maturity and satisfies high standards of reliability and efficiency. After about three periods of development, starting from engineering-oriented methods through mathematical foundations with error analysis and various generalizations to adaptive multiscale and multiphysics simulations of complex micro- and macroprocesses, including for example models for climate and tectonic movements, computational mechanics nowadays is a basic and important subject for teaching and research, and has a multitude of applications.

Last, but definitely not least, the editors would like to thank the team at John Wiley & Sons for their enthusiastic belief in this project, their professionalism, and for creating a friendly atmosphere at the several editorial meetings.

The editors sincerely hope that this encyclopedia will be well accepted by the community and hope that the on-line version with its special features will be used extensively. Of course, we shall be grateful for corrections and proposals for improvements.

Erwin Stein, René de Borst and Thomas J. R. Hughes
Hannover, Delft, Austin
September, 2004

Chapter 1

Fundamentals: Introduction and Survey

Erwin Stein

University of Hannover, Hannover, Germany

| | |
|---|---|
| 1 Motivation and Scope | 1 |
| 2 Stages of Development and Features of Computational Mechanics | 1 |
| 3 Survey of the Chapters of Volume 1 | 2 |
| 4 What We Do Expect | 6 |

1 MOTIVATION AND SCOPE

In the 'Encyclopedia of Computational Mechanics' (ECM), Volume 1 'Fundamentals' includes 26 chapters. It contains the basic methodological, analytical, algorithmic, and implementation topics of computational mechanics.

The main goals of the ECM are to provide first-class up-to-date representations of all major computer-oriented numerical methods and related special features for mechanical problems in space and time, their a priori and a posteriori error analysis as well as various convergent and efficient self-controlling adaptive discretization strategies and to further provide the wide range of robust and efficient direct and iterative solvers as well as challenging applications in all relevant technological areas. Geometrical representations of technical objects, mesh generations, and mesh adaptivity as well as the visualization of input- and output data are also important topics.

The now already 'classical' discretization methods using finite differences, finite elements, finite volumes, and boundary elements were generalized with new conceptual ideas into various directions in the last decade, such as

meshfree methods, spectral, and wavelet techniques as well as discrete finite element algorithms, which are presented here. Error analysis and adaptivity are essential features in general.

2 STAGES OF DEVELOPMENT AND FEATURES OF COMPUTATIONAL MECHANICS

One can say that we are now in about the third period of development of computer-based numerical methods, especially based on weighted residuals, such as the finite element method (FEM) and its generalizations as well as the boundary integral equation method (BIEM or simply BEM) and various couplings of both. The finite difference method (FDM) further plays an important role, especially for time integrations.

The *first period* was from about 1960 to 1975, with a lot of separate engineering approximations for specific mathematical models, especially FEMs for linear elastic static systems, like beams and plates with plane stress states and bending, as well as eigenvalue analysis of stability and vibration problems with applications to structural mechanics. Parallel to this, FEMs for aero- and hydrostatic problems and also for hydrodynamic processes were developed.

The *second period* from about 1976 to 1990 was characterized by rigorous mathematical analysis of Ritz-Galerkin-type discretization methods with trial and test functions in Sobolev spaces within finite subdomains, in order to analyze elliptic boundary-value problems, together with the a priori and a posteriori error analysis of the FEM and the BEM in their various forms for large classes of problems,

such as boundary-value problems of symmetric, positive definite elliptic operators of second order, and also for parabolic and hyperbolic operators with operator splits in space and time, solving systems of ordinary differential equations in time by a finite difference method. Parallel to this, sophisticated engineering developments took place toward complicated linear and nonlinear problems of the classical partial differential equations (PDEs) in mathematical physics with large dimensions of the algebraic equations, motivated and driven by the fast growth of computer power and memory as well as the availability of efficient software systems and, of course, by technological needs and motivations. Numerous chapters in Volumes 2 and 3 show the wide range of challenging applications of FEM, BEM, and various other problem-oriented discretization methods. These new integral methods of weighted residuals are characterized by two important properties:

- (a) *Methodical width:* This is the intended simple logical and algorithmic structure (e.g. with symmetry properties and well-posedness in regular cases) and the possibility of extensions and generalizations within a large class of similar problems, including higher dimensions, and thus forming the frame for a box-type program structure. This is mainly achieved by operations within the finite element subdomains only, that is, without needing neighborhood information on element level, and thus allowing unified assembling and solution procedures of the global systems of algebraic equations. The methods yield relatively small condition numbers of the algebraic equation systems and thus provide robust solutions in regular cases.
- (b) *Methodical depth:* This means the rather simple extension of methods, algorithms, and computer programs to more complicated – especially geometrically and physically nonlinear – problems and to physically coupled problems. This also holds for the implementation of sensitivity analysis within the solution of optimization problems.

These two properties (a) and (b) are the reasons for the tremendous development in and the flexible availability of related program systems for applications in science and technology.

In the third period, from 1991 until now, new tasks, challenges, and research directions can be observed in computational mechanics (more general in applied physics) and in computational mathematics that can be summarized as follows:

- Meshfree and particle methods, finite elements with discontinuities for damage and fracture.

- Error-controlled adaptive modeling and approximation of physical events near to nature, also scale-bridging modeling on different space and timescales, including homogenizations between them.
- Adaptive micromechanical modeling and computation in material science and engineering, including damage, phase changes, and various failure processes.
- New types of generalized FEM and BEM with hierarchical, spectral, and wavelet-based interpolations.
- Modeling and simulation of multiphysics phenomena in science and engineering.
- Complex models and simulations in biomechanics and human medicine.
- New generalized methods for geometrical modeling, mesh generation, and mesh adaptivity.
- New direct and iterative solvers with multilevel and domain decomposition methods.
- Advanced visualization of objects, processes, and numerical results, 3D-animation of virtual reality.

With the advanced current hardware and software tools on hand, about 10 million unknowns of a complex problem can be computed today in reasonable time, using problem-oriented iterative algebraic solvers and preconditioners with advanced data management for high-end machines with parallel or serial scalar and vector processors. Personal computers also enable us to solve hundreds of thousands of unknowns together with error estimation and adaptive mesh refinements. With these tools, it has become possible to realize the verification and even the restricted validation of engineering systems and processes, taking into account disturbed input data and deterministic or statistic imperfections of structures and materials. This leads us to new paradigms in computational mechanics, namely, guaranteeing reliability, safety, and efficiency of results very near to the physical reality of the investigated objects. And because of this progress, computational mechanics helps to simulate virtual products and processes without the necessity of many physical experiments and thus reduces costs and development time of new products considerably.

3 SURVEY OF THE CHAPTERS OF VOLUME 1

Volume 1 can be classified into the four groups: *discretization methods* (Chapter 2 to Chapter 15 of this Volume (14 chapters)); *geometrical modeling, mesh generation, and visualization* (Chapter 16 to Chapter 18 of this Volume (3 chapters)); *Solvers* (Chapter 19 to Chapter 23 of this Volume (5 chapters)); and *time-dependent problems* (Chapter 24 to Chapter 26 of this Volume (3 chapters)).

The first group, **discretization methods**, begins with **Finite difference methods** by Owe Axelsson in which elliptic, parabolic, and hyperbolic problems of second and fourth order as well as convection–diffusion problems are treated in a systematic way, including error analysis and adaptivity, emphasizing computational issues.

Next, FEMs are presented in six chapters, beginning with **Interpolation in finite element spaces** by Thomas Apel, with a survey of different types of test and trial functions, investigating the interpolation error as a basis for a priori and a posteriori error estimates of finite element methods. For a priori estimates, nodal interpolants are used as well as the maximum available regularity of the solution to get optimal error bounds. A posteriori error estimates of the residual type need local interpolation error representations for functions from the Sobolev space $W^{1,2}(\Omega)$. Different interpolation operators and related error estimates are presented for the h -version of the usually used 2D- and 3D-finite elements.

The following chapter, **Finite element methods**, by Susanne Brenner and Carsten Carstensen, treats the displacement method (primal finite element method) for boundary-value problems of second-order elliptic PDE's as well as a priori and a posteriori error estimates of the weak solutions and related h -adaptivity, including non-conforming elements and algorithmic aspects. This basic chapter is followed by **The p -version of the finite element method** for elliptic problems, by Barna Szabó, Alexander Düster, and Ernst Rank, in which hierarchical shape functions of order p are used as test and trial interpolations of the finite elements instead of nodal basis functions. Exponential convergence rates in conjunction with sufficient h -refinement in subdomains with large gradients of the solution are advantageous against the h -version. Boundary layers of dimensionally reduced models (by appropriate kinematic hypotheses), which need the solutions of the expanded mathematical model, can be represented in a consistent way by using adequate p -orders, see also **Chapter 8, this Volume**, by Monique Dauge *et al.* The arising problems are: (i) the fast integration of fully populated element stiffness matrices, (ii) relatively large algebraic systems with strongly populated global stiffness matrices, (iii) the problem of geometric boundary representations without producing artificial singularities, (iv) hp -adaptivity for 3D-systems as well as (v) the efficient implementation of anisotropic p -extensions that are efficient for geometrically and physically anisotropic problems like thin plates and shells, for example, with anisotropic layers of composites. All these problems have been tackled successfully such that the p -type finite element method – in connection with some available computer programs – has reached the necessary maturity for engineering practice.

In **Chapter 6 to Chapter 8 of this Volume**, problem-oriented effective test and trial spaces are introduced for BVPs of PDEs.

Chapter 6, this Volume, by Claudio Canuto and Alfio Quarteroni, is devoted to the high-order trigonometric and orthogonal Jacobi polynomial expansions to be applied to generalized Galerkin methods for periodic and nonperiodic problems, with numerical integration via Gaussian integration points in order to achieve high rates of convergence in total.

Chapter 7, this Volume, by Albert Cohen, Wolfgang Dahmen, and Ronald De Vore, represents matrix compression methods for the BIEM based on wavelet coordinates with application to time-dependent and stationary problems. Wavelets also yield sparsity for the conserved variables of problems with hyperbolic conservation laws. In addition, a new adaptive algorithm is derived for sparse functions and operators of linear and nonlinear problems.

Chapter 8, this Volume, by Monique Dauge, Erwan Faou, and Zohar Yosibash, treats known and new methods for consistent reductions of the 3-D theory of elasticity to 2-D theories of thin-walled plates and shells by expansions with respect to small parameters, without applying the traditional kinematic and static hypotheses. A polynomial representation of the displacements is presumed, depending on the thickness direction, generating singularly perturbed boundary layers in the zero thickness limit. This favors (hierarchical) p -extensions in the thickness direction, yielding hierarchical plate and shell models. Finite element computations show convergence properties and the efficiency of this important problem of boundary layer analysis of plates and shells.

Chapter 9 to Chapter 11 of this Volume treat **Generalized finite element methods**. **Chapter 9, this Volume**, by Ferdinando Auricchio, Franca Brezzi, and Carlo Lovadina, gives a systematic survey and some new results on the stability of saddle-point problems in finite dimensions for some classical mechanical problems, like thermal diffusion, the Stokes equations, and the Lamé equations. Mixed methods yield the mathematical basis for problems with locking and other numerical instability phenomena, like nearly incompressible elastic materials, the Reissner and Mindlin plate equations, and the Helmholtz equation. From an engineering point of view, reduced integration schemes and stabilization techniques get a sound foundation by the problem-dependent *inf-sup* condition. **Chapter 10, this Volume**, by Antonio Huerta, Ted Belytschko, Sonia Fernández-Méndez, and Timon Rabczuk, provides an advanced and systematic representation of different versions and alternatives of the so-called meshfree and particle methods, known as moving least squares, partition of unity FEM, corrected gradient methods, particle-in-cell methods

and so on. The method was originally invented for moving singularities, and discontinuities like crack propagation in solids, in order to avoid frequent complicated and costly remeshings. These methods are based on Ritz-Galerkin- and Petrov-Galerkin-type weighted residual or collocation concepts and generalizations, such as Lagrange multiplier and penalty methods. Radial basis functions are a good tool (without having a compact support) as well as hierarchical enrichments of particles.

The error-controlled approximation of the essential boundary conditions of a boundary-value problem and, of course, the related a priori and a posteriori error analysis as well as the relatively large condition number of the algebraic systems combined with big computational effort are crucial points.

In generally speaking, meshfree methods are now superior or at least equivalent to classical FEMs for some of the addressed specific types of problems.

The last of the three chapters dealing with generalized FEMs, is **Chapter 11, this Volume**, by Nenad J.N. Bicanic. Instead of constructing a convergent and stable numerical method for the approximated solution of, for example, a boundary-value problem for a continuous differential operator, a direct computational simulation of an a priori discrete system with embedded discontinuous deformations, cracks, fragmentations, and so on is treated here. This also includes assemblies of particles of different shapes with their various contact problems, compactions, and other scenarios of real processes. This is a rather new area of computational mechanics, so far mostly treated on an engineering level, that is, without mathematical analysis. The question arises how 'convergence' and 'numerical stability' can be defined and analyzed herein. But there is no doubt that this type of direct computational simulation of technological problems will play an important role in the future.

The two chapters that follow are devoted to **Boundary element methods and their coupling with finite element methods**. **Chapter 12, this Volume**, by George C. Hsiao and Wolfgang L. Wendland, represents variationally based Galerkin-BEMs for elliptic boundary-value problems of second order in a mathematically rigorous way, classified by the Sobolev index. Various boundary integral equations can be derived, introducing fundamental solutions, Green's representation formula, Cauchy data, and four boundary integral operators. Out of this reservoir, several numerical methods and algorithms for boundary elements are presented and discussed. The main features such as stability, consistency, and convergence as well as adequate solvers, condition numbers, and efficiency aspects are well treated. Of course, error analysis and adaptivity play an important role. BEM has advantages over FEM in the case of

complicated boundaries, for example, mechanical problems with edge notches and regular inner domains, and with respect to dimensional reduction by one. Efficient recursive integration formulas and solvers for the fully populated system matrices are available.

Chapter 13, this Volume, by Ernst Stephan, treats the obvious variant of combining the different strengths of both methods by symmetric couplings, which, of course, need considerable algorithmic efforts and adequate solvers with problem-dependent preconditioners. Special features are Signorini-type contact problems using both primal and dual-mixed finite element approximations. Recent features are adaptive hp-methods. There seems to be a lack of available software for 2D- and even more for 3D- problems.

Chapter 14, this Volume, by J. Donea *et al.*, is to be seen separate from the previous presentations of different variational discretization methods, as it treats various coupled processes – for example, fluid-solid interaction – by suitable coordinates and metrics for each of the constituents – for example, Lagrangian coordinates for solids and Eulerian coordinates for fluids – using the well-known tangential push-forward and pull-back mappings between the two descriptions via the deformation gradient. The profit of computational efficiency and robustness can be significant, for example, for the rolling contact of a tire on a street. The ALE-concept, its analysis, the algorithms, and important applications for linear and especially nonlinear static/dynamic problems in solid and fluid mechanics are systematically presented and illustrated by adequate examples. Also, smoothing and adaptive techniques for the finite element meshes are discussed. It is remarkable how quickly the ALE concept was implemented in commercial programs.

Chapter 15, this Volume, by Timothy Barth and Mario Ohlberger, also stands separate from the scheme of finite domain and boundary element approximations. Finite volume elements were invented in fluid mechanics and are also applied now in other branches like biology, and in solid mechanics, too. The advantage of finite volume approximations in comparison with the usual finite element discretizations in finite subdomains is the intrinsic fulfillment of local conservation properties, like mass conservation or entropy growth. Finite volume elements usually also yield robust algebraic systems for unstructured meshes; they are especially favorable for nonlinear hyperbolic conservation systems in which the gradients of the solution functions can blow up in time. Integral conservation laws and discrete volume methods are applied using various meshing techniques of cell- and vertex-centered control volumes. A priori and a posteriori error estimates are presented and

applied for adaptivity. Also, solvers in space and time are discussed.

Chapter 17 to Chapter 18 of this Volume are devoted to the **Computer representation and visualization of topology, geometry, meshes, and computed data**.

Chapter 16, this Volume, by Franz-Erich Wolter, Niklas Peinecke, and Martin Reuter, treats a subject growing in importance in computational mechanics as technical objects and their physical substructures become more and more complicated. The presented methods and realizations can be classified as computational methods for topology and geometry with volume- and boundary-based (direct and indirect) representations of objects and also with a rather new type of modeling, using medial axes and surfaces for describing objects that are mainly one or two dimensional in their appearance. This medial modeling allows a natural transition to finite element meshes. Of course, additional attributes can be included in the geometry, like photometric or toughness properties.

In **Chapter 17, this Volume**, by P.L. George *et al.*, the techniques of planar, surface, and volume meshing as well as adaptive global and local remeshing for discretization methods are outlined, aiming at automatic self-controlled algorithms for various types of mesh generations and their visualizations. Hard problems arise with large 3D-meshes, requiring spatial decomposition, as well as related automatic remeshing and local mesh adaptivity. Moving boundaries and the meshing errors of specific elements with respect to the given analytic or free-form geometry are crucial problems.

The main methods for structured and unstructured mesh generations are presented, where unstructured meshes are constructed in a purely algebraic way or are based on appropriate PDE-solutions. Hierarchical spatial decompositions are used for arbitrary shaped domains with unstructured meshes, like the quadtree and octree method, advancing front strategies and Delaunay type methods.

Chapter 18, this Volume, by William J. Schroeder and Mark S. Shephard, also treats a crucial subject in modern science and technology. Selected interactive visualization of real and virtual objects and processes conveys intrinsic conceiving of the essentials that is hardly possible with data files only. Visualization concerns geometry with attribute data, meshes, and results that may need scalar, vector, and tensor graphics with special features like producing streams of particles or physical quantities through a total system or through a control volume. The presented visualization algorithms give information about what is possible today.

Chapter 19 to Chapter 23 of this Volume treat the crucial problem of stable robust and efficient solvers for the various discrete algebraic systems introduced in the

first 14 chapters. Regarding the mostly high dimensions of algebraic equation systems, which are needed today in science and engineering and which can be solved now in reasonable execution time with the teraflop generation of computers, a variety of sophisticated and problem-oriented types of direct and iterative solvers with adapted preconditioners were developed and are presented here.

Chapter 19, this Volume, by Henk A. van der Vorst, presents direct elimination and iterative solution methods where the latter usually needs efficient preconditioning operators for efficient solutions. All important iterative solvers – based on Krylov projections – are treated, like Conjugate Gradients, MINRES, OMR, Bi-CGSTAB, and GMRES.

Special eigenvalue problems of a Hermitian matrix are analyzed mainly with iterative QR-methods, emphasizing tridiagonal and (upper) Hessenberg matrices. The Krylov subspace approach is an efficient strategy that is applied with four different versions. Several preconditioners are presented and discussed.

In **Chapter 20, this Volume**, by Wolfgang Hackbusch, fast iterative solvers for discretized linear and nonlinear elliptic problems are treated, yielding (optimal) linear complexity of the computational effort in regular cases, which, of course, is of dominant importance to systems with millions of unknowns for which multiprocessor and massively parallel computers are also efficient.

Owing to the smoothing property of the Jacobi or Gauss-Seidel iteration for elliptic PDEs on fine grids, only data of coarse grid points are prolonged after some smoothing steps, and this is repeated through several grids, for example, in a W-cycle with four to five grid levels, such that the solution takes place only with a small equation system. The backward computation with restriction operators and further smoothing steps finishes an iteration cycle. It is efficient to produce a hierarchy of discretizations that is easy for regular and nested grids. Such a hierarchy may also be a side-product of adaptive mesh refinements.

Major issues of the chapter are the complete algorithmic boxes as well as the analysis and error analysis of multigrid methods for FEM and BEM.

Chapter 21, this Volume, by Wolfgang Hackbusch, presents efficient solvers for fully populated matrices as they arise in the boundary integral equation method. The goal is the reduction of $O(n^2)$ arithmetic operations for standard matrix-vector multiplications to nearly $O(n)$ operations. The essential parts of the algorithm are the far-field expansion and the panel cluster tree. A generalized variant is the construction of hierarchical matrices (H-matrices) with different matrix-vector and matrix-matrix

operations only. The computation of inverse stiffness matrices in FEM (e.g. for multiload cases) can be efficiently computed by cluster techniques.

Chapter 22, this Volume, by V.G. Korneev and Ulrich Langer, treats effective iterative solvers for large algebraic systems by the alternating Schwarz method and advanced substructuring techniques, emphasizing efficient problem-dependent preconditioners. Nonoverlapping Schwarz methods are favored against overlapping methods, which need more effort for computer implementation and the control of the solution process. Of course, multiprocessor computers are most suitable for parallel solution within the decomposed domains.

In **Chapter 23, this Volume**, by Werner C. Rheinboldt, strategies for efficient solvers of highly nonlinear algebraic systems, especially with physical instabilities like bifurcation and turning points, are investigated. These solvers are based on Newton's iterative method, its variants, and inexact Newton methods.

The problem of solution instabilities, which depend on one scalar parameter, is analyzed by homotopy methods and continuation methods. Also, the bifurcation behavior of parameterized systems is investigated.

The last three chapters of Volume 1, **Chapter 24, Chapter 25, and Chapter 26 of this Volume**, are devoted to the fundamentals of numerical methods for **Time-dependent problems**.

Chapter 24, this Volume, by Kenneth Eriksson, Claes Johnson, and Anders Logg, is based on duality techniques, by solving associated linearized dual problems for the a posteriori error analysis and adaptivity, using the residuals of Galerkin approximations with shape functions, continuous in space and discontinuous in time, yielding the crucial stability factors and discretization error estimates in adequate norms.

Parabolic initial boundary value problems are usually stiff, and the main problem is the control of error accumulation in time. This is treated successfully with implicit and explicit time-stepping methods for some classical parabolic equations, like the stationary heat equation and the reaction-diffusion problem.

In **Chapter 2, Volume 2**, by Martin Costabel, parabolic and hyperbolic initial boundary value problems with transient solutions in time are considered, like heat con-

duction, diffusion, acoustic scattering, and elastic waves. The following three approaches are critically compared: space-time integral equations, Laplace-transform methods, and time-stepping methods; many advanced mathematical tools are necessary for the analysis, especially the error analysis, which is treated here in a systematic way and illustrated by examples.

Chapter 26, this Volume, by Leszek Demkowicz, treats finite element approximations of the time-harmonic Maxwell equations. With the stabilized variational formulations and Nedelec's three fundamental elements, hp-discretizations and hp-adaptivity are presented. Tetrahedral elements of the first and second type, hexahedral elements of the first type, prismatic elements as well as parametric elements are treated.

The three Nedelec elements deal with the exact sequence of gradient, curl, and divergence operators, yielding the null-space, in addition to projection-based interpolations of finite elements such that the element shape functions are defined as a dual basis to the d.o.f.-functionals, aiming at locality, global continuity, and optimality, and lastly the de Rham commuting diagram property of the analytical and the finite dimensional solution spaces, applying the gradient, curl, and divergence operators.

4 WHAT WE DO EXPECT

We are convinced that the above sketched 26 chapters of Volume 1 are a sound basis for today's and tomorrow's computational mechanics, integrating mathematics, computer science, and physics, especially mechanics, as well as challenging industrial applications that need high computer power. Computer implementations will only be competitive in engineering praxis if they are robust, stable, and efficient, also concerning the requested logical clearness, as well as the width and depth of the algorithms, as explained above.

An important benefit of this encyclopedia is seen in the combination of Volume 1 with Volumes 2 and 3, which are devoted to computational solid and fluid mechanics such that users interested in theoretical issues and those interested in practical issues can both get the information they want, together with any secondary or background knowledge.

Chapter 2 Finite Difference Methods

Owe Axelsson

University of Nijmegen, Nijmegen, The Netherlands

| | |
|---|----|
| 1 Introduction | 7 |
| 2 Two-point Boundary Value Problems | 9 |
| 3 Finite Difference Methods for Elliptic Problems | 12 |
| 4 Finite Difference Methods for Parabolic Problems | 18 |
| 5 Finite Difference Methods for Hyperbolic Problems | 28 |
| 6 Convection-Diffusion Problems | 36 |
| 7 A Summary of Difference Schemes | 50 |
| References | 52 |
| Further Reading | 53 |

1 INTRODUCTION

Although, in general, more restricted in their applicability, finite difference methods provide a simple and readily formulated approximation framework for various types of partial differential equations. They can, hence, often compete with other methods such as finite element methods, which are based on certain variational formulations of the differential equations, and where the preparation work to construct the corresponding systems of algebraic equations is more involved.

When constructing the approximation scheme, it is important to know the type of the given differential equation. Partial differential equations of second order are classified according to the type of their principal

part $\sum_{i,j=1}^n a_{ij}(\mathbf{x})(\partial^2 u / \partial x_i \partial x_j)$. With no limitation, we can assume that the coefficient matrix $A(\mathbf{x}) = [a_{ij}(\mathbf{x})]_{i,j=1}^n$ is symmetric.

As is well known, the differential equation is *elliptic* in \mathbf{x} if all eigenvalues of A have the same sign, *hyperbolic* if one eigenvalue has an opposite sign to the others, and *parabolic* if one eigenvalue is zero and the remaining have the same sign.

We do not consider other cases here. Note that a differential equation can be of different types in different parts of the domain (Ω) of definition. For these classes of differential equations, we need different types of boundary conditions in order for the problem to be well posed.

Definition 1. A boundary value problem is well posed if

- Existence:* There exists a solution that satisfies the equation and each given boundary condition (assuming they are sufficiently smooth).
- Uniqueness:* It has at most one solution.
- Stability:* Its solution is a continuous function of the given data, that is, small changes in the given data entail small changes in the solution.

Although it is of interest in practice to consider also certain ill-posed problems, these will not be dealt with here.

Solutions of partial differential equations of different types have different properties. For instance, the solution to elliptic and parabolic problems at any given point depends on the solution at all other points in Ω . However, for hyperbolic problems, the domain of dependence is a subset of Ω . Although some problems may belong to the class of elliptic problems, they may exhibit a dominating hyperbolic nature in most parts of Ω . To illustrate the aforesaid, consider the following examples.

The problem

$$\mathcal{L}u = -\varepsilon \Delta u + \mathbf{v} \cdot \nabla u = 0 \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega \quad (1)$$

where $|\mathbf{v}| \gg \varepsilon$, $\varepsilon > 0$ has a dominating hyperbolic nature in the interior domain, except near the outflow boundary part, where $\mathbf{v} \cdot \mathbf{n} \geq 0$, and where the diffusion part dominates and the solution has a thin layer. Here \mathbf{n} is the outward-pointing normal vector to Ω .

For the Poisson problem

$$-\Delta u = \rho(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^3$$

where we ignore the influence of boundary data, the solution is

$$u(\mathbf{x}) = \frac{1}{2\pi} \int_{\Omega} |\mathbf{x} - \xi|^{-1} \rho(\xi) d\xi \quad (2)$$

called the *gravitational or electric potential*, where ρ is the density of mass charge. Hence, the solution at any point in Ω depends on all data. On the other hand, the first-order hyperbolic problem

$$au_x + bu_y = 0, \quad x > 0, y > 0 \quad (3)$$

where a, b are positive constants, and with boundary conditions $u(0, y) = g(y)$, $y \geq 0$ and $u(x, 0) = h(x)$, $x \geq 0$ has a solution

$$u(x, y) = \begin{cases} h\left(x - \frac{a}{b}y\right) & x \geq \frac{a}{b}y \\ g\left(y - \frac{b}{a}x\right) & x < \frac{a}{b}y \end{cases}$$

Therefore, the solution is constant along the lines $bx - ay = \text{const}$, called *characteristic lines*, so the solution at any point depends just on a single-boundary data.

Depending on its type, there exist various methods to prove uniqueness and stability of a boundary value problem. For elliptic and parabolic problems, one can use a maximum principle, which shows pointwise stability. As an example, consider $-\Delta u = f$ in Ω , $u = g$ on $\partial\Omega$, where $f \leq 0$. Then, assuming first the contrary, it is readily proven that the maximum of u is taken on the boundary $\partial\Omega$. This also shows that any perturbation of boundary data by some amount ε cannot result in any larger sized perturbation of the solution in the interior of the domain. An alternative way of proving the maximum principle is via a Green's function (fundamental solution) representation of the solution such as in (2).

For hyperbolic problems, one uses the energy integral method. Let u_1 and u_2 be two solutions of $u_t = au_x$, $a > 0$, $0 < x < 1$, $t > 0$, where $u(0, t) = 0$, which correspond to different initial data. Then, $v = u_1 - u_2$ satisfies

the corresponding homogeneous equation. By multiplying the equation by v and integrating, we get

$$\int_0^1 v_t v dx + \int_0^1 av_x v dx = 0$$

or, by letting $E(t) = \int_0^1 v^2 dx$ (called the *energy integral*), we find

$$\frac{1}{2} \frac{d}{dt} E(t) + \frac{a}{2} v^2(1, t) = 0, \quad t > 0$$

that is, $E'(t) \leq 0$. Hence, $E(t) \leq E(0)$, $t > 0$, where $E(0) = \int_0^1 (u_1(x, 0) - u_2(x, 0))^2 dx$, which shows stability and uniqueness.

Clearly, the energy method is also applicable for more general problems, where the analytical solution is not known. For the discretized problems, one can use a discrete maximum principle, enabling error estimates in maximum norm. For parabolic and hyperbolic problems, we can also use stability estimates based on eigenvalues of the corresponding difference matrix or, more generally, based on energy type estimates. Finally, for constant coefficient problems, Fourier methods can be used.

Let the domain of definition of the continuous problem be discretized by a rectangular mesh with a mesh size h . In order to compute a numerical solution of a partial differential equation like (1), we replace the partial derivatives in the differential equation by finite differences. In this way, we obtain a discrete operator equation

$$L_h u_h = f_h$$

We will be concerned with the well-posedness of the discrete equation, using a discrete maximum principle, for instance, as well as with estimates of the *discretization error* $e_h = u - u_h$, where u is the solution of the original continuous partial differential equation. This estimation will be based on the *truncation error* $\tau_h = L_h u - f_h$.

The following finite difference approximations will be used throughout this text:

$$u'(x) \approx D_x^+ u(x) := \frac{1}{h} [u(x+h) - u(x)], \quad \text{the forward difference} \quad (4)$$

$$u'(x) \approx D_x^- u(x) := \frac{1}{h} [u(x) - u(x-h)], \quad \text{the backward difference} \quad (5)$$

$$u'(x) \approx D_x^0 u(x) := \frac{1}{2h} [u(x+h) - u(x-h)], \quad \text{the first-order central difference} \quad (6)$$

Note that $D_x^0 u = (1/2)(D_x^+ + D_x^-)u$. More generally, one can use the so-called " θ -method"

$$u'(x) \approx [\theta D_x^+ + (1-\theta)D_x^-]u(x) \quad (7)$$

where θ is a method parameter. Thus,

$$u'(x) = [\theta D_x^+ + (1-\theta)D_x^-]u(x) + \frac{h}{2}(1-2\theta)u''(x) - \frac{h^2}{6}[(1-\theta)u^{(3)}(x-\eta_2 h) + \theta u^{(3)}(x+\eta_1 h)]$$

where $0 < \eta_i < 1$, $i = 1, 2$, so

$$u'(x) = [\theta D_x^+ + (1-\theta)D_x^-]u(x) + \begin{cases} O(h^2) & \text{if } |1-2\theta| = O(h) \\ O(h) & \text{otherwise} \end{cases}$$

If $0 \leq \theta \leq 1$, we get

$$u'(x) = \frac{1}{h} [\theta u(x+h) + (1-2\theta)u(x) - (1-\theta)u(x-h)] + \frac{h}{2}(1-2\theta)u''(x) - \frac{h^2}{6}u^{(3)}(x+\eta_3 h), \quad -1 < \eta_3 < 1$$

Note that for $\theta = 1/2$, we get (6), for $\theta = 1$, we get (4), and for $\theta = 0$, we get (5). Hence, the θ -method generalizes the previous methods.

An approximation for the second derivative is obtained as

$$u''(x) \approx D_{xx}^+ u(x) \equiv D_{xx}^0 u(x), \quad \text{the central difference of second order} \quad (8)$$

We have then $D_{xx}^0 u(x) = D_x^+ D_x^- u(x) = h^{-2} [u(x+h) - 2u(x) + u(x-h)]$, thus

$$u''(x) = D_{xx}^0 u(x) - \frac{h^2}{12} u^{(4)}(x+\eta_4 h), \quad -1 < \eta_4 < 1$$

or $u''(x) = D_{xx}^0 u(x) + O(h^2)$, $h \rightarrow 0$

Similar expressions hold for D_{xx}^+ , D_{xx}^- , and so on. In particular, if $u^{(4)}, u^{(6)} \in C(\bar{\Omega})$,

$$\begin{aligned} D_{xx}^+ D_x^- u(x, y) + D_x^+ D_{xx}^- u(x, y) \\ = h_x^{-2} [u(x+h_x, y) + u(x-h_x, y) - 2u(x, y)] \\ + h_y^{-2} [u(x, y+h_y) + u(x, y-h_y) - 2u(x, y)] \\ = u_{xx}(x, y) + u_{yy}(x, y) + O(h_x^2) + O(h_y^2), \quad h_x, h_y \rightarrow 0 \end{aligned} \quad (9)$$

For $h_x = h_y = h$, we have

$$\Delta^{(5)} u := [D_x^+ D_x^- + D_y^+ D_y^-]u(x, y) = h^{-2} [u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h) - 4u(x, y)]$$

$\Delta^{(5)}$ is called the *5-point difference operator*.

Various difference methods adjusted to the type of the problem, with discretization error estimates based on truncation errors, are presented.

When deriving truncation error estimates for difference approximations, one uses Taylor expansion (assuming sufficient regularity of u)

$$u(x_i + h) = u(x_i) + hu'(x_i) + \cdots + \frac{1}{(k-1)!} \times h^{k-1} u^{(k-1)}(x_i) + R(x_i, h, k) \quad (10)$$

where the remainder term $R(x_i, h, k)$ can be written as $R(x_i, h, k) = (1/k!)h^k u^{(k)}(\xi_i)$, $\xi_i \in (x_i, x_i + h)$ or in the alternative form $R(x_i, h, k) = \int_{x_i}^{x_i+h} [1/(k-1)!](x_i + h - s)^{k-1} u^{(k)}(s) ds$.

2 TWO-POINT BOUNDARY VALUE PROBLEMS

The most common among problems of applied mathematics type that appear in physics, engineering, and so on are boundary value problems for partial differential equations. As an introduction to difference methods for such problems, we consider here the corresponding problem in one dimension, the two-point linear differential equation problem:

Find $u \in C^2[a, b]$ such that

$$\mathcal{L}u = -(k(x)u')' + p(x)u = f(x), \quad a < x < b \quad (11)$$

with boundary conditions

$$\begin{aligned} r_0(u) &\equiv \gamma_0 u(a) - \delta_0 k(a)u'(a) = \alpha \\ r_1(u) &\equiv \gamma_1 u(b) + \delta_1 k(b)u'(b) = \beta \end{aligned} \quad (12)$$

Here, $u' = du/dx$, $k(x) \geq k_0 > 0$, $a \leq x \leq b$ and $k \in C^1(a, b)$, and $p, f \in C(a, b)$ are given real-valued functions and $\gamma_0, \delta_0, \gamma_1, \delta_1, \alpha, \beta$ are given real numbers. The operator \mathcal{L} is self-adjoint, that is, $\int_a^b \mathcal{L}u v dx = \int_a^b \mathcal{L}v u dx$. The solution u will then be a twice continuously differentiable function.

Such problems arise, for instance, if we let u be the displacement of a (thin) elastic string subjected to forces with distribution defined by f . In the simplest model, k is constant, $p(x) \equiv 0$, and the string is fixed at (a, α) ,

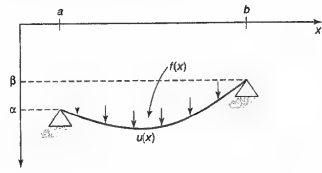


Figure 1. An elastic string subjected to forces.

(b, β), see Figure 1. This equation also arises if we let u be the temperature in a heat-conducting rod that is perfectly insulated except at the endpoints, where the temperatures α, β are given. The material coefficients k and p are often assumed to be constants.

Existence and uniqueness

For boundary value problems, there are intrinsic questions regarding existence and uniqueness of solutions. There holds the following theorem.

Theorem 1 (Fredholm's alternative) Equation (11) has exactly one solution if and only if the homogeneous problem (with $f \equiv 0$ and homogeneous boundary conditions $\alpha = \beta = 0$) only has the trivial solution.

We now give conditions under which the boundary value problem (11) has exactly one solution.

Theorem 2. If $k(x) > 0$, $p(x) \geq 0$, $\gamma_1^2 + \delta_1^2 > 0$, $\delta_i \geq 0$, $\gamma_i \geq 0$, $i = 0, 1$ and if any of the following conditions holds, (i) $\gamma_0 \neq 0$, (ii) $\gamma_1 \neq 0$, (iii) $p(x) > 0$ for some $x \in (a, b)$, then the boundary value problem (11) has exactly one solution.

That the homogeneous problem $-(ku')' + p(x)u = 0$, $a < x < b$, $r_0(u) = r_1(u) = 0$, where the boundary operators r_0, r_1 are defined in (12), has only the trivial solution, can be shown by partial integration of $Lu = 0$.

2.1 Difference approximations

The boundary value problem (11) cannot, in general, be solved analytically. We shall now present a simple but efficient numerical method. In order to simplify the presentation, we shall mostly consider the following

problem:

Find $u \in C^2[a, b]$ such that

$$Lu = -u'' + p(x)u = f(x), \quad a < x < b \quad (13)$$

$$u(a) = \alpha, \quad u(b) = \beta$$

where $p(x) \geq 0$, $a < x < b$, and $\max(\alpha, \beta) \geq 0$.

Let $x_0 = a$, $x_i = x_{i-1} + h$, $i = 1, 2, \dots, N$, $x_{N+1} = b$, where $h = (b-a)/(N+1)$ is a uniform partitioning $\pi = \pi_h$ of the interval $[a, b]$. In order to find an approximate solution at the points x_i , we approximate u'' by finite differences

$$-u''(x_i) \approx h^{-2}[-u(x_{i-1}) + 2u(x_i) - u(x_{i+1})]$$

Letting u_h be the resulting approximation of u at x_i , $i = 1, 2, \dots, N$, we get

$$L_h u_h(x_i) = h^{-2}[-u_h(x_{i-1}) + 2u_h(x_i) - u_h(x_{i+1})] + p(x_i)u_h(x_i) = f(x_i),$$

$$i = 1, 2, \dots, N, \quad u_h(a) = \alpha, \quad u_h(b) = \beta \quad (14)$$

This is a linear system of N equations in the N unknowns $u_h(x_i)$, $i = 1, 2, \dots, N$. That the corresponding matrix is nonsingular follows from a discrete maximum principle.

Lemma 1 (Discrete maximum principle) Let L_h be defined by (14) and v_h be any function with $\max\{v_h(x_0), v_h(x_{N+1})\} \geq 0$ defined on x_i for which $L_h v_h(x_i) \leq 0$, $i = 1, 2, \dots, N$. Then,

$$\max_{0 \leq i \leq N+1} v_h(x_i) = \max_{i=0, N+1} v_h(x_i)$$

that is, the largest value of v_h is taken at one of the endpoints.

Next, we show the corresponding error estimates for v_h and e_h . We now define the truncation error of the difference approximation as the error (or defect) we get when we substitute the difference approximation u_h in (14) for the exact solution u of (13).

Definition 2. The function $\tau_{h,i} = (L_h u)_i - f_i$, $i = 1, \dots, N$ is referred to as the truncation error $\tau_{h,i}$ of the difference approximation (14) at x_i .

Definition 3. The function $e_h(x_i) = u(x_i) - u_h(x_i)$, $i = 0, 1, \dots, N+1$ is referred to as the discretization error of the difference approximation (14) at point x_i .

Note that the truncation error and the discretization error are related by

$$L_h e_h = \tau_h \quad (15)$$

It turns out that τ_h is easily estimated via Taylor expansions. We see from (15) that in order to estimate e_h , we need a bound of the inverse of the discrete operator L_h (if it exists). Such a bound is provided by the next barrier function lemma, which can be proven by the discrete maximum principle.

In order to estimate e_h , it is convenient to use the following notations:

$$|v_h|_{\pi_h} = \max_{1 \leq i \leq N} |v_h(x_i)|, \quad |v_h|_{\partial \pi_h} = \max_{i=0, N+1} |v_h(x_i)| \quad (16)$$

Lemma 2 (Barrier function lemma) Assume that there exists a function w_h (called barrier function), defined on x_i , $i = 0, 1, \dots, N+1$, which satisfies $L_h w_h > 0$, $w_h \geq 0$ and $w_h(x_0) = 0$. Then, any function v_h defined on x_i , $i = 0, 1, \dots, N+1$, for which $v_h(x_0) = 0$ satisfies

$$|v_h|_{\pi_h \cup \partial \pi_h} \leq |v_h|_{\partial \pi_h} + \frac{\max_{x_i} w_h}{\min_{x_i} L_h w_h} |L_h v_h|_{\pi_h} \quad (17)$$

Lemma 3. (a) $\tau_{h,i} = h^{-2}[-u(x_{i-1}) + 2u(x_i) - u(x_{i+1})] + u''(x_i)$.

(b) If $u \in C^{(4)}(a, b)$, then

$$\tau_{h,i} = -\frac{1}{12}h^2 u^{(4)}(\xi_i), \quad x_{i-1} < \xi_i < x_{i+1}$$

Proof. Since $f_i = f(x_i) = p(x_i)u(x_i) - u''(x_i)$ and $(L_h u)_i = h^{-2}[-u(x_{i-1}) + 2u(x_i) - u(x_{i+1})] + p(x_i)u(x_i)$, (a) follows. (b) follows now by Taylor expansion around x_i . Hence, if $u \in C^{(4)}(a, b)$, the truncation error is $O(h^2)$, $h \rightarrow 0$. \square

In order to prove that the discretization error is also $O(h^2)$, we use Lemma 2. As a barrier function, we use

$$w_h(x_i) = 1 - \left(\frac{b+a-2x_i}{b-a}\right)^2 \quad (18)$$

Note that $w_h(a) = w_h(b) = 0$.

Theorem 3. The discretization error of the difference approximation (14) satisfies

$$|u - u_h|_{\pi_h \cup \partial \pi_h} \leq \frac{1}{96}[(b-a)h]^2 \max_{a \leq \xi \leq b} |u^{(4)}(\xi)| \quad (19)$$

Proof. Note that w_h , defined by (18), satisfies $1 \geq w_h \geq 0$ on π_h and, as is easily seen, $L_h w_h = 8/(b-a)^2 + p(x)w_h(x_i) \geq 8/(b-a)^2 > 0$. By Lemmas 2 and 3(b), we now get

$$|u - u_h|_{\pi_h \cup \partial \pi_h} \leq |u - u_h|_{\partial \pi_h} + \frac{(b-a)^2}{8} |L_h(u - u_h)|_{\pi_h}$$

$$= \frac{(b-a)^2}{8} |L_h u - f|_{\pi_h} = \frac{(b-a)^2}{8} |\tau_h|_{\pi_h}$$

$$\leq \frac{(b-a)^2}{8} \frac{h^2}{12} \max_{a \leq \xi \leq b} |u^{(4)}(\xi)|.$$

\square

2.2 Richardson extrapolation

For a uniform mesh and sufficiently smooth solutions, the order of accuracy of the approximate solution can be easily improved using the classical trick of Richardson extrapolation and nested meshes involving approximations at same meshpoints for two different values of h .

Theorem 4. Let $u \in C^6(a, b)$ and let u_h be the solution of (14). Then,

$$u(x_i) - u_h(x_i) = h^2 \varphi(x_i) + O(h^4),$$

$$h \rightarrow 0, \quad i = 1, 2, \dots, N$$

where φ is a function that is independent of h .

Proof. Letting φ be the solution of the auxiliary problem

$$-\varphi'' + p\varphi = -\frac{1}{12}u^{(4)}, \quad a < x < b, \quad \varphi(a) = \varphi(b) = 0$$

and assuming that $\varphi \in C^4(a, b)$, which holds if $u \in C^6(a, b)$, it follows that $L_h(e - h^2\varphi) = O(h^4)$, $h \rightarrow 0$, where $e = u - u_h$. Applying the Barrier Lemma completes the proof. \square

This argument can be repeated so that if u is smooth enough, one can prove an h^2 expansion of the error $u(x_i) - u_h(x_i)$. Hence, repeated Richardson extrapolation may also be applied. Such error expansions do not, however, always exist. For instance, if one of the boundary points is not a meshpoint in the uniform mesh, the distance δh , $0 < \delta < 1$ of this boundary point to the nearest meshpoint is not a smooth function of h .

The extrapolation procedure has advantages over conventional higher-order methods. Thus, the basis difference formula can be very simple, which makes repetition with a new mesh size easy and the method automatically provides error estimates.

For extensions of such asymptotic error expansions and Richardson extrapolation for linear finite elements; see Blum, Lin and Rannacher (1986).

2.3 Computing high-order approximations of the derivatives of the solution

We now show that not only the solution but also its derivatives can be computed to high accuracy, using the already computed approximate values of the solution.

Theorem 5. Let $u \in C^6(a, b)$ be the solution of $-u'' + p(x)u = f$, $a < x < b$, $u(a) = \alpha$, $u(b) = \beta$ and let u_h be the solution of the discrete problem, discretized using central differences on a uniform mesh. Then,

$$u'(x_i) = \frac{u_h(x_{i+1}) - u_h(x_{i-1}))}{2h} + O(h^2)$$

Proof. Theorem 4 shows that $u_h(x_i) = u(x_i) - h^2\varphi(x_i) + O(h^4)$, where φ does not depend on h and $\varphi \in C^2(a, b)$. Hence, using Taylor series expansions,

$$\begin{aligned} \frac{u_h(x_{i+1}) - u_h(x_{i-1}))}{2h} &= \frac{u(x_{i+1}) - u(x_{i-1}))}{2h} \\ &\quad - h^2 \frac{\varphi(x_{i+1}) - \varphi(x_{i-1}))}{2h} + O(h^3) = u'(x_i) \\ &\quad + \frac{h^2}{6} u^{(3)}(\eta_1) - h^2 \varphi'(x_i) + O(h^3) \end{aligned}$$

where $\eta_1 \in (x_{i-1}, x_{i+1})$. \square

In a similar way, assuming a correspondingly higher order of regularity of the solution, even higher-order derivatives can be computed with error $O(h^2)$. This result is not obvious, because to compute an approximation of u' , we make use of approximations of u , divided by h or a higher power of h . That we do not lose one or more power(s) of h in the order of approximation is due to the existence of an h -expansion of the errors.

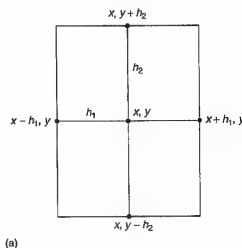
3 FINITE DIFFERENCE METHODS FOR ELLIPTIC PROBLEMS

We present in this section various difference methods for the numerical solution of partial differential equations of elliptic type. Discretization errors are derived for operators of positive type. The derivations are done for problems in two space dimensions but most results can be readily extended to problems in three space dimensions. Further

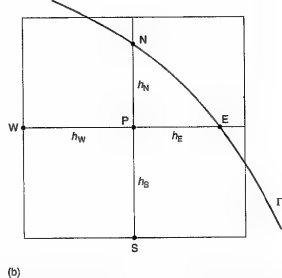
results on difference methods for elliptic problems can be found in Section 6.

3.1 Difference approximations

Consider first the Poisson problem $\Delta u = f$ with given Dirichlet boundary conditions on a unit square domain, aligned with the coordinate system, and a uniform rectangular mesh Ω_h with mesh sizes $h_1 = \rho_1 h$ and $h_2 = \rho_2 h$ in the x - and y -direction respectively. Here, ρ_1 and ρ_2 are given positive numbers, chosen such that $1/\rho_1 h$ is an integer, and h is a small positive parameter, which we intend to make arbitrarily small to get a sufficiently accurate numerical solution. In each interior meshpoint in Ω_h (see Figure 2a), we use the five-point difference approximation,



(a)



(b)

Figure 2. Local difference meshes (a) rectangular domain, (b) curved boundary.

$\Delta_h^{(5)} u := (D_{h_1}^+ D_{h_1}^- + D_{h_2}^+ D_{h_2}^-) u(x, y)$ and on the boundary points, we use the given boundary values.

Let $h_i = 1/(N_i + 1)$, $i = 1, 2$ and, hence, let $\rho_1/\rho_2 = (N_1 + 1)/(N_2 + 1)$, where $N_i + 1$, $i = 1, 2$ are the number of meshpoints in the two coordinate directions. Let u_h denote the corresponding approximate solution. We then get a system of $N_1 N_2$ linear algebraic equations of the form

$$\begin{aligned} L_h u_h &:= h_1^{-2} [u_h(x-h_1, y) - 2u_h(x, y) + u_h(x+h_1, y)] \\ &\quad + h_2^{-2} [u_h(x, y-h_2) - 2u_h(x, y) + u_h(x, y+h_2)] \\ &= f(x, y), \quad x, y \in \Omega_h \end{aligned} \quad (20)$$

If we order the meshpoints in lexicographic (horizontal) order, the corresponding matrix takes the form

$$A_h = \text{block_tridiag}(I_1, A^{(0)}, I_1)$$

where $A^{(0)} = h_1^{-2} \text{tridiag}(1, -2, 1) - 2h_2^{-2} I$ and $I_i = h_2^{-2} I$, $i = 1, 2, \dots, N_2$. Here, $A^{(0)}$ and the identity matrix I have order $N_1 \times N_1$ and there are N_2 block-rows in A_h . Systems with matrix A_h can be solved efficiently using various methods such as fast direct methods or multigrid and algebraic multilevel iteration methods; see, for example, Axelsson and Barker (1984) and Hackbusch (1985).

Assuming that $u \in C^4(\Omega)$, the local truncation error $L_h u - f$ of the difference approximation is readily found to be

$$\begin{aligned} L_h u - f &= \frac{1}{12} h_1^2 u_{xx}^{(4)}(x + \xi_1, y) + \frac{1}{12} h_2^2 u_{yy}^{(4)}(x, y + \xi_2) \\ &= O(h^2), \quad -h_i < \xi_i < h_i, \quad i = 1, 2 \end{aligned}$$

Curved boundaries

For a more general domain Ω , for instance with a curved boundary such as illustrated in Figure 2b, we must modify the approximation scheme at interior points in Ω_h next to the boundary. There are two such efficient schemes. The first uses a generalized five-point difference approximation with, in general, nonquidistant meshlengths (see Shortley and Weller, 1938):

$$\begin{aligned} L_h u_h &:= \left\{ \frac{1}{h_E} [u_h(E) - u_h(P)] - \frac{1}{h_W} [u_h(W) - u_h(P)] \right\} \\ &\quad \times \frac{2}{h_W + h_E} + \left\{ \frac{1}{h_N} [u_h(N) - u_h(P)] \right. \\ &\quad \left. - \frac{1}{h_S} [u_h(S) - u_h(P)] \right\} \frac{2}{h_S + h_N} = f(P), \quad P \in \Omega_h \end{aligned} \quad (21)$$

where h_E, h_W, h_N, h_S denote the distances from P to the surrounding points in the East, West, North, and South

directions, see Figure 2. Unless $h_E = h_W$ and $h_N = h_S$, the local truncation error is $O(h)$ here. The coefficient matrix is in general not symmetric. The second method uses a linear combination of weighted linear interpolations in the x - and y -directions,

$$\begin{aligned} L_h u_h &:= \frac{1}{h_E} [h_E u_h(W) + h_W u_h(E) - (h_E + h_W) u_h(P)] \\ &\quad + \frac{1}{h_N} [h_N u_h(S) + h_S u_h(N) - (h_N + h_S) u_h(P)] \\ &= \frac{1}{2} \left[h_W \frac{h_E + h_W}{2} + h_S \frac{h_S + h_N}{2} \right] f(P), \quad P \in \Omega_h \end{aligned} \quad (22)$$

Here, the coefficient matrix is symmetric.

Remark 1. To provide an alternative to the Shortley-Weller approximation for treating curved boundaries, much work has been devoted to numerical grid generation during the past 20 to 30 years. For instance, curvilinear boundary-fitted finite difference methods became popular and applied extensively in numerical fluid dynamics problems (see Thompson, Warsi and Mastin, 1985). More recently, much effort has been devoted to variational grid generation methods, which can provide more robust methods applicable also for very complicated geometries; see, for example, Garanzha (2004) and the references there in.

3.2 Higher-order schemes

3.2.1 The nine-point difference scheme

Let $\Delta u = f$, and consider first the cross-directed five-point difference scheme on a local equidistant square submesh,

$$\begin{aligned} \Delta_h^{(5, \times)} &:= \frac{1}{2} h^{-2} [u_h(x-h, y-h) + u_h(x+h, y-h) \\ &\quad + u_h(x-h, y+h) + u_h(x+h, y+h) \\ &\quad - 4u_h(x, y)], \quad x, y \in \Omega_h \end{aligned} \quad (23)$$

It is readily seen that, for a sufficiently smooth function u ,

$$\begin{aligned} \Delta_h^{(5)} &= \Delta u + \frac{2}{4!} h^2 (u_{xx}^{(4)} + u_{yy}^{(4)}) + \frac{2}{6!} h^4 (u_{xx}^{(6)} + u_{yy}^{(6)}) \\ &\quad + O(h^6) \\ \Delta_h^{(5, \times)} &= \Delta u + \frac{2}{4!} h^2 (u_{xx}^{(4)} + 6u_{xxyy} + u_{yy}^{(4)}) + \frac{2}{6!} h^4 \\ &\quad \times (u_{xx}^{(6)} + 15u_{xxxxyy} + 15u_{xyxyyy} + u_{yy}^{(6)}) + O(h^6) \end{aligned}$$

Let $\Delta_h^{(9)}$ be the nine-point difference scheme defined by

$$\Delta_h^{(9)} = \frac{2}{3} \Delta_h^{(5)} + \frac{1}{3} \Delta_h^{(5, \times)}$$

The coefficients in this stencil equal $1/6$ for the corner vertex points in the square with edges $2h$, equal $2/3$ for the midedge points, and equal $-10/3$ for the center point.

A computation shows that for a uniform rectangular mesh,

$$\Delta_h^{(9)} u_h = f + \frac{1}{12} h^2 \Delta f + \frac{1}{360} h^4 (\Delta^2 f + 2f_{xxyy}) + O(h^6)$$

where $\Delta^2 = \Delta(\Delta f)$. Using a modified right-hand side in the difference formula, it follows that the difference approximation

$$\Delta_h^{(9)} u_h = \left[I + \frac{h^2}{12} \Delta_h^{(9)} \right] f, \quad (x, y) \in \Omega_h \quad (24)$$

has truncation error $O(h^4)$.

Further, it follows from (24) that for a sufficiently smooth function f , $\Delta f = \Delta_h^{(9)} f - (1/12)h^2 \Delta^2 f + O(h^4)$. A computation shows that $h^2 f_{xxyy} = 2[\Delta_h^{(5, \times)} f - \Delta_h^{(9)} f] + O(h^4)$ and, therefore, the nine-point stencil with the next modified right-hand side

$$\Delta_h^{(9)} u_h = f + \frac{1}{12} h^2 \Delta_h^{(9)} f - \frac{1}{240} h^4 \Delta_h^{(5)} \Delta f + \frac{1}{180} h^4 f_{xxyy}, \quad (x, y) \in \Omega_h$$

has a truncation error $O(h^6)$.

The implementation of this scheme is simplified if f is given analytically so that Δf and so on can be computed explicitly. If $f = 0$, then $\Delta_h^{(9)} u_h = 0$ has an order of approximation $O(h^6)$. Hence, this scheme provides very accurate approximation, for instance, for far-field equations, where frequently $\Delta u = 0$.

The above is an example of a compact difference scheme (for further references on such schemes, see Houstis and Rice, 1982).

3.2.2 Difference methods for anisotropic problems and problems with a mixed derivative

Consider first the anisotropic differential equation $au_{xx} + bu_{yy} = f(x, y)$, $(x, y) \in \Omega$, where $u = g(x, y)$, $(x, y) \in \partial\Omega$ and f and g are given, sufficiently smooth functions. Let $a > 0$ and $b > 0$.

Here, the nine-point difference approximation has a stencil, as shown in (25) with $c = 0$. If we modify the right-hand side to be $f + 1/12 h^2 \Delta f$, it can be seen that in this case the local truncation error becomes $O(h^4)$.

Consider next the differential equation with a mixed derivative

$$au_{xx} + 2cu_{xy} + bu_{yy} = f(x, y), \quad (x, y) \in \Omega$$

with given boundary conditions. We assume that $a > 0$, $b > 0$, and $c^2 < ab$, which are the conditions for ellipticity of the operator. For the mixed derivative, we use the central difference approximation $u_{xy} \approx D_x^0 D_y^0 u$, that is,

$$u_{xy} \approx \frac{1}{4h^2} [u_h(x-h, y-h) - u_h(x+h, y-h) - u_h(x-h, y+h) + u_h(x+h, y+h)]$$

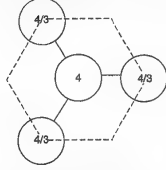
Combined with the nine-point difference stencil, it becomes

$$\frac{1}{6} h^{-2} \begin{bmatrix} a+b-3c & 5b-a & a+b-3c \\ 5a-b & -10(a+b) & 5a-b \\ a+b-3c & 5b-a & a+b-3c \end{bmatrix} \begin{bmatrix} u_h(x-h, y-h) \\ u_h(x, y-h) \\ u_h(x+h, y-h) \\ u_h(x-h, y) \\ u_h(x, y) \\ u_h(x+h, y) \\ u_h(x-h, y+h) \\ u_h(x, y+h) \\ u_h(x+h, y+h) \end{bmatrix} = f_h(x, y) \quad (25)$$

3.2.3 Difference schemes for other regular tessellations

Finite differences can be extended to nonrectangular meshes.

For a regular (isosceles) triangular mesh, one can form the obvious seven-point difference stencil. For a hexagonal ('honeycomb') mesh, one finds a four-point stencil



The symmetrically located nodepoints in the seven-point scheme allows one to readily approximate second-order cross derivatives. Similarly, in 3D problems, a cuboctahedral stencil involves 13 nodepoints. If a Cartesian grid is used, approximations of the second-order cross derivatives require at least nine points in 2D and 19 points in 3D, that is, two and six more than for the hexagonal and cuboctahedral stencils.

The biharmonic operator $\Delta^2 u = f$, $(x, y) \in \Omega$ with boundary conditions such as $u = g(x, y)$, $\partial u / \partial n = q(x, y)$ on $\partial\Omega$ give rise to a 12-point stencil for a regular equidistant mesh,

$$h^{-4} \begin{bmatrix} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{bmatrix} u_h = f_h \quad (26)$$

which has truncation error $O(h^2)$. The biharmonic problem can, however, more efficiently be solved as a coupled problem

$$\begin{cases} \xi - \Delta u = 0 \\ \Delta \xi = f \end{cases}$$

using a variational formulation hereof; see, for example, Axelsson and Gustafsson (1979) and Ciarlet and Raviart (1974).

3.3 Approximating boundary conditions

So far, we have only considered Dirichlet type boundary conditions $u = g$ on $\partial\Omega$. For a Neumann ($\partial u / \partial n := \nabla u \cdot n = g$) or the more general Robin ($\partial u / \partial n + \sigma u = g$) (cf. Gustafson and Abe, 1995) type boundary conditions, one must use special approximation methods. The regular difference mesh is then extended around the boundary line. If the normal to the boundary goes through meshpoints, we can use a central difference approximation for $\partial u / \partial n$, using an interior meshpoint and its symmetrically reflected point in the extended domain (see Figure 3(a)). In other cases, we can still use central difference approximations for $\partial u / \partial n$ (at U, R in Figure 3(b)) but we must interpolate the function value in the symmetrically located points in the interior. This can be done using linear interpolation from the surrounding points (P, N, NW, W) in Figure 3(b) to find the value at T , or using biquadratic interpolation, involving some additional points. The local truncation error becomes $O(h)$ or $O(h^2)$, respectively. It can be seen that one can always get a positive-type scheme in the first case, but not in the second case.

For discretization errors for problems with curved boundaries, see Section 6.

3.4 Monotonicity and discretization error estimates

Monotone operators provide a basis for pointwise bounds of the solution and the discretization errors corresponding to various difference approximations.

The general form of a linear difference operator L_h depending on some discretization parameter h is

$$L_h u_h(P_i) = \sum_{j=1}^n a_{ij}(P_i) u_h(P_j) = \tilde{f}(P_i), \quad i = 1, 2, \dots, n$$

Here, the function \tilde{f} includes the given source function and the given boundary data.

The operator is said to be monotone if $L_h v \geq 0$ implies $v \geq 0$, where v is a function defined on Ω_h . Note that

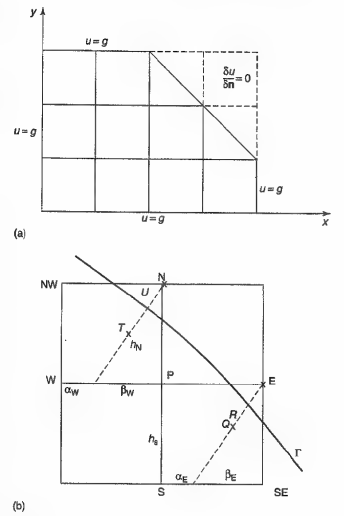


Figure 3. Neumann boundary conditions.

a monotone operator is nonsingular because if L_h is monotone and $L_h v \leq 0$, then $L_h(-v) \geq 0$, so $-v \geq 0$, that is, $v \leq 0$. Hence, if $L_h v = 0$, then both $v \geq 0$ and $v \leq 0$ hold, so $v = 0$.

It is also readily seen that a nonsingular matrix A is monotone if and only if $A^{-1} \geq 0$, where the inequality holds componentwise. Further, for any monotone operator A , there exists a positive vector $v > 0$ such that $Av > 0$. Take, for example, $v = A^{-1}e$, where $e = [1, 1, \dots, 1]^T$. If some component $v_i = 0$, it would follow that all entries of the i th row of A^{-1} were zero, which is clearly not possible.

Consider now a monotone difference operator L_h and a normalized function w with $\max_i w(P_i) = 1$, called a barrier function, such that $L_h w(x) \geq \alpha$ for all $x \in \Omega_h$ and for some positive constant α . Then,

$$\|L_h^{-1}\|_{\infty} \leq \frac{1}{\alpha}$$

where $\|L_h\|_\infty = \sup_{v \neq 0} \|L_h v\|_\infty / \|v\|_\infty$ and $\|v\|_\infty = \max_i |v(x_i)|$, $x_i \in \Omega_h$ is the supremum norm.

As is readily seen, this yields a discretization error estimate,

$$\|e_h\|_\infty = \|u - u_h\|_\infty \leq \frac{1}{\alpha} \|L_h u - f_h\|_\infty$$

where u is the exact solution to a given differential equation $Lu = f$, L_h is a discrete approximation to L , u_h is the solution to the corresponding discrete equation $L_h u_h = f_h$, and $L_h u - f_h = L_h u - L_h u_h$ is the truncation error.

The barrier function thus leads to an easily computable discretization error estimate if the discrete operator is monotone. In addition, as pointed out by Collatz (1986), the monotonicity property is of great practical value because it can readily be used to get lower and upper bounds for the solution.

Monotone operators arise naturally for positive-type difference schemes. The difference operator L_h is of positive type if the coefficients satisfy the sign pattern $a_{ij} > 0$ and $a_{ij} \leq 0$, $j \neq i$.

If, in addition, $a_{ij} \geq \sum_{j=1}^n |a_{ij}|$, $i = 1, 2, \dots, n$, $j \neq i$ with strong inequality for at least one index i , then the operator is monotone. This is equivalent with that the matrix $A = [a_{ij}]_{i,j=1}^n$ is an M -matrix.

Stieltjes proved that a positive definite operator of positive type is monotone.

However, even if the operator is not of positive type, it might nevertheless be monotone. For instance, a familiar result states that if $A = M - N$ is a weak regular splitting (i.e. M is monotone and $M^{-1}N \geq 0$), then A is monotone if and only if the splitting is convergent, namely, there holds $\rho(M^{-1}N) < 1$ with ρ being the spectral radius of $M^{-1}N$.

Hence, monotonicity of a given matrix A , respectively, a linear discrete operator L_h , can be proven by constructing convergent weak regular splittings $A = M - N$. As is shown below, this result can be extended to matrix splittings of a more general form.

3.4.1 Bounding inverses of monotone matrices

To bound the supremum norm of the inverse of a nonsingular monotone matrix A , we consider the following Barrier Lemma both in its general and sharp forms.

Lemma 4 (The Barrier Lemma) Let A be a monotone matrix of order n and let v be a normalized vector, $\|v\|_\infty = 1$, such that $\min_i (Av)_i \geq \alpha$ for some positive scalar α . Then,

$$(a) \|A^{-1}\|_\infty \leq 1/\alpha;$$

$$(b) \|A^{-1}\|_\infty = 1/\max\{\min_i (Av)_i, v \in V_A\}, \text{ where } V_A = \{v \in \mathbb{R}^n, \|v\|_\infty = 1, Av > 0\};$$

$$(c) \|A^{-1}\|_\infty = \|x\|_\infty \text{ where } x \text{ is the solution of } Ax = e.$$

Proof. For a proof, see Axelsson and Kolotilina (1990). \square

For later use, note that for a strictly diagonally dominant matrix $A = [a_{ij}]$, it holds the inequality

$$\|A^{-1}\|_\infty \leq \frac{1}{\min_i \left\{ |a_{ii}| - \sum_{j \neq i} |a_{ij}| \right\}}$$

We now extend the barrier lemma to the case where the positive vector v satisfies the weaker condition $Av \geq 0$. This result can be particularly useful if Dirichlet boundary conditions hold at some part of the domain Ω .

Lemma 5. Let A be monotone of the form $A = \begin{bmatrix} A_{11} & -A_{12} \\ -A_{21} & A_{22} \end{bmatrix}$, where A_{11} is monotone, A_{12} and A_{21} are nonnegative and A_{22} has no zero rows. Further, let $v = [v_1; v_2]^T$ be a positive vector such that $\|v\|_\infty = 1$ and $Av_1 \geq 0$, $v_2 > 0$, $q \geq 0$. Then there holds

$$\|A^{-1}\|_\infty \leq \left(1 + \frac{\|A_{12}\|_\infty}{\min_i (A_{11}v_1)_i} \right) \left(1 + \frac{\|A_{21}\|_\infty}{\min_i (A_{11}v_1)_i} \right) \times \max \left\{ \frac{1}{\min_i (A_{11}v_1)_i}, \frac{1}{\min_i (q + A_{21}v_1)_i} \right\} \quad (27)$$

A proof can be found in Axelsson and Kolotilina (1990).

3.4.2 Proving matrix monotonicity

Here, we summarize some classical results on weak regular splittings, convergent splittings, and Schur complements of matrices partitioned into a two-by-two block form, which can be used to ascertain that a given matrix is monotone.

Let $M, N \in \mathbb{R}^{n \times n}$; then, $A = M - N$ is called a weak regular splitting if M is monotone and $M^{-1}N$ is nonnegative. The splitting is convergent if $\rho(M^{-1}N) < 1$.

Theorem 6. A weak regular splitting $PA = M - N$, where P is nonsingular and nonnegative, is convergent if and only if A is monotone.

Proof. See Axelsson and Kolotilina (1990). \square

In practical applications, it can be more convenient to use, instead of Theorem 6, the following sufficient conditions.

Theorem 7. Let $PAQ = M - N$ be a weak regular splitting with P and Q nonsingular and nonnegative. Then, A is monotone if there exists a positive vector v such that either $M^{-1}PAQv > 0$ or $v^T M^{-1}PAQ > 0$.

Proof. Since by assumption $M - N$ is a weak regular splitting, it follows by Theorem 6 that PAQ is nonsingular and monotone if $\rho(M^{-1}N) < 1$. But $M^{-1}PAQv = (I - M^{-1}N)v$ or, since $M^{-1}N$ is nonsingular, $0 \leq M^{-1}Nv = (I - M^{-1}PAQ)v < v$ if $M^{-1}PAQv > 0$. Hence, with $D = \text{diag}(v_1, v_2, \dots, v_n)$, that is, $De = v$, $0 \leq D^{-1}M^{-1}ND e < e$ or $\|D^{-1}M^{-1}ND\|_\infty < 1$, so $\rho(M^{-1}N) < 1$.

In a similar way, if $v^T M^{-1}PAQ > 0$, it follows that $\|DM^{-1}ND^{-1}\|_1 < 1$, where $\|\cdot\|_1$ is the ℓ_1 -norm. \square

Remark 2. Theorem 7 can be particularly useful when we have scaled A with diagonal matrices P and Q .

From Theorem 7, one can deduce the following important monotonicity comparison condition.

Corollary 1. Let $B_1 \leq A \leq B_2$, where B_1 and B_2 are monotone matrices. Then, A is monotone and $B_2^{-1} \leq A^{-1} \leq B_1^{-1}$.

Proof. See Axelsson and Kolotilina (1990). \square

Theorem 8. Let $A = [A_{ij}]$ be an $m \times m$ block matrix satisfying the following properties

- (i) A_{ij} are nonsingular and $A_{ii}^{-1} > 0$, $i = 1, 2, \dots, m$.
- (ii) For $i \neq j$, there exist matrices $P_{ij} \geq 0$ such that $A_{ij} \leq -P_{ij}A_{jj}$.
- (iii) There exists a positive vector v such that either the block components $(A^T v)_i$ are nonzero and nonnegative for $i = 1, 2, \dots, m$ or $Av > 0$, where $u_i = A_{ii}^{-1}v_i$, $i = 1, 2, \dots, m$.

Then A is monotone.

Proof. Let $D = \text{blockdiag}(A_{11}, \dots, A_{mm})$. By (i), $A_{ii}^{-1} > 0$ and by (ii), $A_{ij}A_{jj}^{-1} \leq 0$, $i \neq j$, implying that $AD^{-1} \leq I$. Now Theorem 7 with $P = M = I$, $Q = D^{-1}$ can be applied to prove monotonicity of A if $v^T AD^{-1} > 0$ or $AD^{-1}v > 0$, which, however, holds by (iii). (Note that we have assumed strict inequality, $A_{ii}^{-1} > 0$.) \square

The following theorem shows that monotonicity of two-by-two block matrices holds if and only if its Schur complement is monotone.

Theorem 9. Let A be a two-by-two block matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} and A_{22} are square submatrices and A_{11} is nonsingular.

- (a) If A_{11} is monotone and $A_{11}^{-1}A_{12}$ and $A_{21}A_{11}^{-1}$ are nonnegative, then A is monotone if $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ is monotone.
- (b) Conversely, assume that A is monotone. Then S is monotone.

Proof. To prove the existence of the inverse in part (a), use the block matrix factorization of A , which shows that A is invertible if and only if A_{11} and S are invertible.

The monotonicity statements in both parts follow from the explicit form of the inverse of A ,

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix}$$

Example 1. Consider the mixed derivative elliptic problem

$$-au_{xx} - 2cu_{xy} - bu_{yy} = f \text{ in } \Omega = [0, 1]^2 \quad (28)$$

$u = 0$ on $\partial\Omega$, with variable coefficients $a(x, y)$, $b(x, y) > 0$, and $c(x, y) \geq 0$. After elimination of boundary conditions, the standard nine-point second-order accurate finite difference approximation of this problem on a uniform mesh, yields the block tridiagonal $n^2 \times n^2$ matrix

$$A = \frac{1}{h^2} \text{block-tridiag} \left[T \left(-\frac{c_i}{2}, -b_i, \frac{c_i}{2} \right), T(-a_i, 2(a_i + b_i), -a_i), T \left(-\frac{c_i}{2}, -b_i, -\frac{c_i}{2} \right) \right]$$

where $T(a_i, b_i, c_i)$ stands for a tridiagonal matrix with diagonal coefficients b_i and off-diagonal a_i and c_i . Let

$$B = \frac{1}{h^2} \text{block-tridiag} \left[T \left(0, -b_i, \frac{c_i}{2} \right), T(-a_i, 2(a_i + b_i), -a_i), T \left(\frac{c_i}{2}, -b_i, 0 \right) \right]$$

Clearly, $A \leq B$ and by Theorem 7 with $P = Q = I$, $M = B$, monotonicity of A follows if the inequality $B^{-1}Ae > 0$ and monotonicity of B hold.

The diagonal blocks of B are clearly monotone, and since the block components $(Be)_i$, $i = 1, 2, \dots, n$ are nonzero and nonnegative, applying Corollary 1 to B^T , we conclude

that B is monotone if

$$c_i \leq \frac{a_i b_{i+1}}{a_{i+1} + b_{i+1}}, \quad c_i \leq \frac{a_i b_{i-1}}{a_{i-1} + b_{i-1}}$$

By a symmetry argument, B is also monotone if

$$c_i \leq \frac{a_{i-1} b_i}{a_{i-1} + b_{i-1}}, \quad c_i \leq \frac{a_{i+1} b_i}{a_{i+1} + b_{i+1}}$$

To prove the monotonicity of A , it remains to be shown that $B^{-1}Ae > 0$. Clearly, $(Ae)_i$ is nonzero and nonnegative for $i = 1, 2, \dots, n$, and since the diagonal blocks of B^{-1} are positive and $B^{-1} \geq 0$, the required inequality follows.

Note that in the constant coefficient case, the conditions above take the form

$$c \leq \frac{ab}{a+b} \quad (29)$$

which is stronger than the ellipticity condition $c < (\sqrt{ab})$. Note also that in the matrix in (25), where a nine-point stencil has been used for the approximation of $au_{xx} + bu_{yy}$, the difference stencil is of positive type only if $|c| \leq (a+b)/6$.

4 FINITE DIFFERENCE METHODS FOR PARABOLIC PROBLEMS

In this section, we discuss the numerical solution of parabolic problems of the form

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad x \in \Omega, \quad t > 0 \quad (30)$$

with initial condition $u(x, 0) = u_0(x)$ and given boundary conditions on $\partial\Omega$, valid for all $t > 0$. Here, L is an elliptic operator and f is a given, sufficiently smooth function. The boundary conditions can be of general, Robin type, $(\partial u / \partial n) + \sigma(u - g(x, t)) = 0$, where $\sigma \geq 0$, g is given and n is the outer unit vector normal to $\partial\Omega$. Here, $\sigma = \infty$ corresponds to Dirichlet boundary conditions. Frequently, in applications we have $L = \Delta$, the Laplace operator. As is well known, this equation is a model for the temperature distribution in the body Ω , for instance. The equation is called the *heat conduction* or *diffusion* equation.

Stability and uniqueness of the solution of problem (30) can be shown using a maximum principle, which holds for nonpositive f , or decay of energy for the homogeneous problem. Such and other properties of the solution can be important for the evaluation of the numerical solution methods.

The equation can be solved by a semidiscretization method, such as the *method of lines* as it has also been called. In such a method, one usually begins with discretization of the space derivatives in the equation, which leaves us with a system of ordinary differential equations in variable t , that is, an initial value problem. The system is stiff, and to enable choosing the time-steps solely based on approximation properties one uses stable implicit methods. In particular, a simple method called the θ -method can be used.

Alternatively, we may begin with discretization in time using, for instance, the θ -method. This results in an elliptic boundary value problem to be solved at each time-step, which can be done using the methods presented in Section 3.

Usually, the order in which we perform the discretizations, first in space and then in time, or vice versa, is irrelevant in the respect that the same algebraic equations, and hence the same numerical solution, result at each time-step. However, the analysis of the methods may differ. Also, if we intend to use a variable (adaptive) mesh in space, it is more natural to begin with the discretization in time. At various time-steps, we can then use different approximations in space.

4.1 Properties of the solution

4.1.1 A maximum principle

For ease of exposition, we describe the maximum principle for an equation of the form (30) in one space variable. On the other hand, we allow for a domain whose boundary may depend on time. (Such problems arise in so-called free boundary value problems, where the boundary between two matters, such as ice and water, may vary with time. Frequently, the temperature of ice is assumed to be constant and it remains to compute the temperature distribution in the water. Such a problem also arises in connection with permafrost.)

Hence, let the domain D be defined by the indicated parts of the boundary lines

$$L_0 = \{(x, 0) \mid \phi_1(0) \leq x \leq \phi_2(0)\}$$

$$L_1 = \{(x, T) \mid \phi_1(T) \leq x \leq \phi_2(T)\}, \quad T > 0$$

and the curves

$$K_1 = \{(\phi_1(t), t) \mid 0 \leq t \leq T\}$$

$$K_2 = \{(\phi_2(t), t) \mid 0 \leq t \leq T\}$$

where $\phi_1(t) < \phi_2(t)$ are continuous functions. Let $\Gamma_0 = L_0 \cup K_1 \cup K_2$ and $\Gamma = \Gamma_0 \cup L_1$ (see Figure 4).

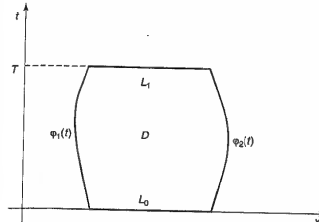


Figure 4. Domain of definition for a parabolic problem.

Theorem 10. If u is the solution of

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad (x, t) \in D \quad (31)$$

where $f \leq 0$, and if u is sufficiently smooth in D , then $\max_{(x,t) \in D} u(x, t) \leq \max_{(x,t) \in \Gamma_0} u(x, t)$, that is, u takes its maximum value on Γ_0 .

Proof. By contradiction. \square

The same maximum principle can be proven for problem (30). From this maximum principle, uniqueness and stability of solutions of (30) follow.

Theorem 11. Let $D = \Omega \times [0, T]$, $t > 0$.

- Problem (30) has at most one solution in D , which takes prescribed boundary and initial conditions on Γ_0 .
- If the boundary and initial conditions are perturbed by an amount of at most ϵ , then the solution is perturbed by at most this amount.
- If f in (30) is perturbed by a nonpositive function (but the boundary and initial conditions are unchanged), the solution u is also perturbed by a nonpositive function.

As an application of the maximum principle, we consider the derivation of two-sided bounds of the solution.

Corollary 2. Let $Ku = u'(t) + Lu = f(t)$, $t > 0$ and let \underline{u} and \bar{u} be two sufficiently smooth functions that satisfy $K\underline{u} \leq f \leq K\bar{u}$ in D , $\underline{u} \leq \bar{u}$ on Γ_0 . Then $\underline{u} \leq u \leq \bar{u}$.

4.1.2 Exponential decay of energy

Consider a heat equation with a reaction term

$$u_t = u_{xx} + au, \quad 0 < x < 1, \quad t > 0 \quad (32)$$

where $u(0, t) = u(1, t) = 0$ and $u(x, 0) = u_0(x)$ is a given sufficiently smooth function. We assume that a is a constant, satisfying $a \leq K - c$, where $K = (\|u_x\|/\|u\|)^2$, c is positive constant and $\|u\|^2 = \int_0^1 u^2 dx$. Letting $E(t) = (1/2) \int_0^1 u^2(x, t) dx$ (the square L_2 norm, a measure of energy) and using (32), we find that $\int_0^1 u_x u dx = - \int_0^1 u_x^2 dx + a \int_0^1 u^2 dx$, that is, $E'(t) = \|u\|^2(a - K) \leq -c\|u\|^2 = -2cE(t)$, or

$$E(t) \leq e^{-2ct} E(0), \quad t > 0$$

with $E(0) = (1/2) \int_0^1 u_0^2 dx$. Hence, $E(t) \rightarrow 0$ exponentially, as $t \rightarrow \infty$.

The constant K can take arbitrary large values. For example, for $u(x, t) = \sin k\pi x g(t)$ and $g \neq 0$, $K = (k\pi)^2$, and here k can be arbitrary large. By the classical Sobolev inequality, there holds that $K \geq \pi^2$. Hence, if $a \leq 0$, then we can take any $c \leq K$, or $c = \pi^2$ and $E(t) \leq e^{-2\pi^2 t} E(0)$, $t > 0$.

A similar result holds also for more general parabolic problems $u_t + Lu = au$, where the operator L is coercive, for example, $\int_\Omega Lu dx \geq \rho \int_\Omega u^2 dx$ for some positive constant ρ .

4.1.3 Exponential decay of the solution

The solution of a parabolic problem depends on initial data on the whole space Ω at all times. However, an exponential decay holds not only for the energy but also for the solution, away from the support of the initial function, assuming it has a bounded support. This observation can be based on the explicit form of the solution of the pure initial value problem (referred to as the Cauchy problem),

$$u_t = u_{xx} \text{ in } -\infty < x < \infty, \quad t > 0$$

$$u(x, t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(x-y)^2}{4t}} u_0(y) dy \quad (33)$$

where $u(x, 0) = u_0(x)$, $u_0 = 0$ outside a bounded domain Ω . From (33), it readily follows that

$$|u(x, t)| = O\left(|x|^{-1} e^{-\frac{x^2}{4t}}\right) \text{ as } |x| \rightarrow \infty$$

which shows a rapid decay away from the region of compact support of the initial function. Formula (33) also shows that the solution $u(x, t)$ is an infinitely differentiable function of x and t for any positive t and that a similar decay holds for its derivatives.

Remark 3. Since, by a partitioning of unity, any initial data can be partitioned in packets of initial data with compact support, it can hence be efficient for linear problems

to compute the solution for each such data packet, even in parallel, at least for the smallest values of t . Since the domain of influence widens with $O(\sqrt{t})$, as t increases, we must, however, add an increasing number of contributions to the solution from several such subdomains to get the solution at each point in the solution domain.

4.2 Finite difference schemes: the method of lines

For the numerical solution of parabolic problems, one commonly uses a semidiscretization method.

Consider first a discretization of space derivatives in (30), for instance by use of central difference approximations. Then, for each nodepoint in the mesh $\Omega_h \subset \bar{\Omega}$, including the boundary mesh points, where the solution u is not prescribed, we get an approximation $U(t)$, which is a vector function of t , and whose i th component approximates u at $x_i \in \Omega_h$ at time t . The vector function U satisfies the system of ordinary differential equations we get when the operator \mathcal{L} in (30) is replaced by a matrix A_h , corresponding to a difference approximation of Δ . Hence,

$$\frac{dU(t)}{dt} = A_h U(t) + b(t), \quad t > 0 \quad (34)$$

where $b(t)$ contains the values of $f(x, t)$ at $x = x_i$ and any nonhomogeneous boundary condition. For $t = 0$, we have that the i th component of $U(0)$ satisfies $U_i(0) = u_0(x_i)$ at the meshpoints x_i of Ω_h .

In general, for elliptic operators, A_h has a complete eigenvector space with eigenvalues with negative real parts (see Section 3), and (34) is stable if b is bounded. This method has been called *method of lines* because we approximate u along each line perpendicular to Ω in the time-space domain and beginning in x_i .

4.2.1 Stability of the method of lines

We comment first on the stability of linear systems of ordinary differential equations,

$$\frac{du}{dt} = Au + b(t), \quad t > 0, \quad u(0) = u_0 \quad (35)$$

where A is a $n \times n$ matrix. Its solution is

$$u(t) = \exp(tA)u_0 + \int_0^t \exp[(t-s)A]b(s)ds, \quad t > 0$$

The analysis of stability of such systems can be based on the Jordan canonical form of A ; see, for example, [1] (1996). Without going into full details, we consider only

the case where the homogeneous system is *asymptotically stable*, that is,

$$\|\exp(tA)\| \xrightarrow{t \rightarrow \infty} 0$$

Let $\alpha(A) = \max_i \operatorname{Re} \lambda_i(A)$ be the so-called *spectral abscissa* of A . Analysis shows that the system is asymptotically stable for any perturbation of initial data if and only if $\alpha(A) < 0$. Similarly, for the solution of (35), there holds $\|u(t)\| \rightarrow 0$ if $\alpha(A) < 0$ and $\|b(t)\| \rightarrow 0$ and $\|u(t)\|$ is bounded if $\alpha(A) < 0$ and $\|b(t)\|$ is bounded.

Similar to the energy estimates, an alternative analysis can be based on the spectral abscissa $\beta(A)$ of the symmetric part of A , that is, $\beta(A) = \max_i \lambda_i[(1/2)(A + A^T)]$. We assume that $\beta(A) < 0$.

Then, let $E(t) = (1/2)\|u(t)\|^2$. It follows from (35) that $E'(t) = (Au, u) + (b(t), u) = \frac{1}{2}[(A + A^T)u, u] + (b(t), u)$ where $(u, v) = u^T v$. Hence,

$$E'(t) \leq \beta(A)E(t) + \frac{1}{2}\beta(A)\|E(t) + \beta(A)\|^{-1}\|b(t)\|^2$$

or

$$E'(t) \leq \frac{1}{2}\beta(A)E(t) + \beta(A)\|E(t) + \beta(A)\|^{-1}\|b(t)\|^2$$

that is

$$E(t) \leq e^{\frac{1}{2}\beta(A)t} E(0) + \int_0^t |\beta(A)|^{-1} e^{\frac{1}{2}\beta(A)(t-s)} \|b(s)\|^2 ds$$

and, since $\beta(A) < 0$, if $\|b(t)\| \rightarrow 0$, $t \rightarrow \infty$ then exponential decay of energy follows.

The following relation between $\alpha(A)$ and $\beta(A)$ holds.

Lemma 6 (Dahlquist (1959), Lozinskij (1958)) $-\beta(-A) \leq \alpha(A) \leq \beta(A)$.

Proof. If $Ax = \lambda x$, where $\operatorname{Re}(\lambda) = \alpha(A)$ and $\|x\| = 1$, then $x^* A^* = \bar{\lambda} x^*$, and hence $(1/2)x^*(A^* + A)x = (1/2)(\bar{\lambda} + \lambda) = \alpha(A)$. But, $\beta(A) = \max_{\|x\|=1} (1/2)x^*(A^* + A)x$, by the Rayleigh quotient theory. Hence $\alpha(A) \leq \beta(A)$. Similarly, $\alpha(-A) \leq \beta(-A)$. But $\alpha(A) \geq -\alpha(-A)$, as an elementary consideration shows. Hence $\alpha(A) \geq -\beta(-A)$. \square

Unlike the scalar case, $\sup_{t \geq 0} \|\exp(tA)\|$ may be strictly greater than 1 even though $\alpha(A)$ is negative. The next relation illustrates further how $\|\exp(tA)\|$ may grow and shows exactly when $\sup_{t \geq 0} \|\exp(tA)\| = 1$.

Theorem 12

- (a) $e^{\alpha(A)t} \leq \|\exp(tA)\| \leq e^{\beta(A)t}$, $t \geq 0$.
(b) $\sup_{t \geq 0} \|\exp(tA)\| = 1 \Leftrightarrow \beta(A) \leq 0$.

Proof. Let v be an arbitrary unit vector, $\|v\| = 1$, and let $\phi(t) = v^* \exp(tA^*) \exp(tA) v$. Then,

$$\phi'(t) = (\exp(tA)v)^*(A^* + A)\exp(tA)v \quad (36)$$

that is, by Rayleigh quotient theory, $\phi'(t) \leq 2\beta(A)\phi(t)$. Hence, $\phi(t) \leq e^{2\beta(A)t}$ and for every t ,

$$\sup_{\|v\|=1} \phi(t) = \|\exp(tA)\|^2 \leq e^{2\beta(A)t}$$

This implies the rightmost inequality of (a). To prove the leftmost inequality, note that for any matrix B , $\|B\| \geq \max_i |\lambda_i(B)|$.

Hence,

$$\|\exp(tA)\| \geq \max_{\|v\|=1} |e^{\lambda_i t}| = \max_i e^{\operatorname{Re} \lambda_i t} = e^{\alpha(A)t}$$

which proves (a). By (a), the sufficiency part of (b) is already proven. The converse follows by

$$\sup_{t \geq 0} \|\exp(tA)\| = 1 \Leftrightarrow \phi(t) \leq 1, \quad t \geq 0 \quad \forall v \in C^n, \quad \|v\| = 1$$

Hence, since $\phi(0) = 1$, we have $\phi'(0) \leq 0 \forall v$, $\|v\| = 1$ or by (36), $v^*(A^* + A)v \leq 0 \forall v$, $\|v\| = 1$. Hence, $\beta(A) \leq 0$. \square

Corollary 3. If A is a normal matrix (i.e. $A^*A = AA^*$) that is, A is diagonalizable, then $\beta(A) = \alpha(A)$ and the inequalities in Theorem 12 (a) are sharp: $\|\exp(tA)\| = e^{\alpha(A)t}$.

4.3 The θ -method

For the numerical solution of the system of ordinary differential equations, arising from the method of lines, there exist a plethora of methods. We consider here in more detail only one such method, the so-called θ -method. It is presented for linear problems but it is also readily applicable for certain nonlinear problems.

For (34), the θ -method takes the following form

$$[I - (1 - \theta)kA]V(t + k) = (I + \theta kA)V(t) + k[(1 - \theta)b(t + k) + \theta b(t)] \quad (37)$$

$t = 0, k, 2k, \dots$, where $V(0) = U_0$, I is the identity operator and $V(t)$ is the corresponding approximation of $U(t)$. Here, θ is a method parameter.

The θ -method takes familiar forms for particular values of θ . For example, $\theta = 1$ yields the familiar Euler forward method

$$V(t + k) = V(t) + k[AV(t) + b(t)]$$

$\theta = 0$ yields the backward Euler (or Laasonen) method

$$V(t + k) = V(t) + k[AV(t + k) + b(t + k)]$$

while $\theta = (1/2)$ determines the trapezoidal (or Crank-Nicolson) rule

$$V(t + k) = V(t) + \frac{k}{2}[A[V(t) + V(t + k)] + b(t) + b(t + k)]$$

We see from (37) that the θ -method is *explicit* only for $\theta = 1$; for all other values of θ , the method is *implicit*, requiring the solution of a linear algebraic system of equations at each step. The extra computational labor caused by implicit methods have to be accepted, however, for reasons of stability.

Observe further that $I - (1 - \theta)kA$ is nonsingular if k is small enough or if $\operatorname{Re} \lambda_i \leq 0 \forall i$ and $\theta \leq 1$, where λ_i is an eigenvalue of A .

For the stability analysis of time-stepping methods, one can use the Jordan canonical form of the corresponding system matrix. For a general linear system of first-order difference equations

$$V(t + k) = BV(t) + c(t), \quad t = 0, k, 2k, \dots, \quad V(0) = U_0 \quad (38)$$

It shows that the homogeneous system is asymptotically stable, that is, $\|B^k\| \rightarrow 0$, $k \rightarrow \infty$ if and only if $\mu(B) < 1$, where $\mu(B) = \max_i |\lambda_i(B)|$. Further, the solutions of the inhomogeneous system (38) are bounded if $\mu(B) < 1$ and $\|c(k)\|$ is bounded for $k \rightarrow \infty$ and satisfy $\|x(k)\| \rightarrow 0$, $k \rightarrow \infty$ if $\mu(B) < 1$ and $\|c(k)\| \rightarrow 0$, $k \rightarrow \infty$.

However, even if $\mu(B) < 1$, for nonnormal (i.e. nondiagonalizable) matrices, it may happen that $\|B^k\|$ takes huge values for finite values of k . Hence, in practice, we may have to require that $\|B\| < 1$.

An alternative approach is to use the *numerical radius* $r(B)$ of B , where $r(B) = \max\{x^* B x\}; x \in C^n, (x, x) = 1\}$. It is seen that $r(B) \leq \|B\|$, and it can be shown (see e.g. Axelsson, 1994) that $\|B\| \leq 2r(B)$. It can further be shown that if $r(B) \leq 1$, then $r(B^k) \leq r(B)^k \leq 1$, $k = 1, \dots$ for any square matrix B . Hence, in general, the stronger stability condition $\|B\| < 1$ can be replaced by $r(B) \leq 1$. Unfortunately, the computation of the numerical radius is, in general, complicated. In the case when B is nonnegative, it can be shown that $r(B) = \rho((1/2)(B + B^T))$, which can be used to compute $r(B)$.

4.3.1 Preservation of stability

We show now the conditions when the θ -method preserves the stability of the system (35), that is, for which it holds

$\alpha(A) < 0 \Rightarrow \mu(B) < 1$. We then first put (37) in the form of (38). Thus, (38) holds with

$$B = [I - (1 - \theta)kA]^{-1}[I + \theta kA]$$

$$c(t) = k[I - (1 - \theta)kA]^{-1}[(1 - \theta)b(t + k) + \theta b(t)]$$

Let (λ_j, v_j) , $j = 1, 2, \dots, n$, denote the eigensolutions of A with the ordering

$$Re(\lambda_1) \geq Re(\lambda_2) \geq \dots \geq Re(\lambda_n)$$

Then, the eigensolutions of B are (μ_j, v_j) , $j = 1, 2, \dots, n$, where

$$\mu_j = \frac{1 + \theta k \lambda_j}{1 - (1 - \theta)k \lambda_j} \quad (39)$$

Theorem 13. Assume that (35) is asymptotically stable ($Re(\lambda_1) < 0$) and that $b(t)$ is bounded, all $t \geq 0$. Then

- (a) (38), with $0 \leq \theta \leq (1/2)$, is asymptotically stable $\forall k > 0$. (Unconditional stability.)
 (b) (38), with $(1/2) < \theta \leq 1$, is asymptotically stable if and only if

$$k < -\frac{2Re(\lambda_j)}{(2\theta - 1)|\lambda_j|^2}, \quad j = 1, 2, \dots, \quad (\text{Conditional stability}) \quad (40)$$

Proof. An easy calculation establishes that

$$|\mu_j|^2 = \frac{1 + 2\theta k Re(\lambda_j) + \theta^2 k^2 |\lambda_j|^2}{1 - 2(1 - \theta)k Re(\lambda_j) + (1 - \theta)^2 k^2 |\lambda_j|^2}$$

where μ_j is given by (39). Hence,

$$|\mu_j|^2 < 1 \Leftrightarrow k(2\theta - 1)|\lambda_j|^2 < -2Re(\lambda_j) \quad (41)$$

The assumption that (35) is asymptotically stable, implies that $Re(\lambda_j) < 0$ for $j = 1, 2, \dots, n$. It is easy to see that the inequality on the right side of (41) is satisfied for $j = 1, 2, \dots, n$, precisely under the conditions on k presented above. \square

The solution of the homogeneous system $u'(t) = Au(t)$, $t > 0$, $u(0) = u_0$, where A has a complete eigenvector space, can be written in the form

$$u(t) = \sum_{j=1}^n c_j e^{\lambda_j t} v_j, \quad t > 0$$

where c_1, \dots, c_n are determined by the expansion $u_0 = \sum_{j=1}^n c_j v_j$ and (λ_j, v_j) are the eigenpairs of A . The corresponding solution of the difference equation is

$$V(t) = \sum_{j=1}^n c_j \mu_j^k v_j, \quad t = 0, k, 2k, \dots$$

where $\mu_j = \mu(k\lambda_j; \theta)$ and where we have introduced the function

$$\mu(\beta; \theta) = \frac{1 + \theta\beta}{1 - (1 - \theta)\beta}, \quad -\infty < \beta < \infty$$

Hence, μ_j is the damping factor corresponding to $e^{\lambda_j t}$ (note $Re(\lambda_j) < 0$), and it is important that each factor is sufficiently small, even for the large eigenvalues. There holds

$$\lim_{\beta \rightarrow -\infty} \mu(\beta; \theta) = \begin{cases} -\frac{\theta}{1 - \theta}, & 0 \leq \theta < 1 \\ -\infty, & \theta = 1 \end{cases}$$

Hence, $|\mu_j|$ is less than one only if $\theta < 1/2$. In particular, for $\theta = 1/2$, it holds that $\mu \rightarrow -1$, as $\beta \rightarrow -\infty$. Therefore, the Crank-Nicolson method can have an insufficient damping, especially at the initial, transient, phase (small t) where all eigenvector components may have a significant value. It is then better to choose $\theta < 1/2$, for instance, $\theta = (1/2) - \zeta k$ for some $\zeta > 0$. (The latter choice will preserve the second order of discretization error, as is shown below.) Otherwise, for a fixed value of $\theta < 1/2$, the time-discretization error is only $O(k)$.

4.3.2 Discretization error estimates for the θ -method

The discretization error estimates for the θ -method can be readily derived. The fully discretized scheme takes the form

$$[I - k(1 - \theta)L_h]u(x, t + k) = [I + k\theta L_h]u(x, t) - k(1 - \theta) \times f(x, t - k) + k\theta f(x, t), \quad t = 0, k, 2k, \dots \quad (42)$$

We have $f = u_t - Lu$ and

$$u(x, t + k) - u(x, t) = \int_t^{t+k} u_t(x, s) ds$$

Substituting the differential equation solution in (42), we then obtain that

$$[I - k(1 - \theta)L_h][u(x, t + k) - [I + k\theta L_h]u(x, t) + k(1 - \theta) \times f(x, t + k) + k\theta f(x, t) + \int_t^{t+k} u_t(x, s) ds - k(1 - \theta)$$

$$\times L_h u(x, t + k) - k\theta L_h u(x, t)] - k(1 - \theta)[u_t(x, t + k) - Lu(x, t + k)] - k\theta[u_t(x, t) - Lu(x, t)]$$

Letting $e(x, t) = u(x, t) - v(x, t)$, we then find

$$[I - k(1 - \theta)L_h]e(x, t + k) = [I + k\theta L_h]e(x, t) + \tau(x, t; h, k)$$

where τ is the truncation error

$$\tau(x, t; h, k) = \int_t^{t+k} u_t(x, s) ds - [k(1 - \theta)u_t(x, t + k) + k\theta u_t(x, t)] + k(1 - \theta)(L - L_h)u(x, t + k) + k\theta(L - L_h)u(x, t)$$

Note that the truncation error consists of two parts: (1) the time-discretization error and (2) the space-discretization error. The stability analysis shows that they can be estimated independently of each other and the discretization error in L_2 norm in space becomes

$$\|u(\cdot, t) - v(\cdot, t)\| \leq [(\theta - \frac{1}{2})O(k)] \sup_{t>0} \|u_{tt}\| + O(k^2) \sup_{t>0} \|u_{ttt}\| + O(h^2)(\|u^{(2)}\| + \|u^{(3)}\|), \quad h, k \rightarrow 0 \quad (43)$$

If we choose $\theta = (1/2) - \zeta k$ for some positive constant ζ , then the total discretization error becomes $O(k^2) + O(h^2)$. Full details of such an analysis and applications for nonlinear problems of parabolic type can be found in Axelsson (1984).

Remark 4. Note that we were able to derive this estimate without the use of the monotonicity of $(-L_h)$ (cf. Section 3). Hence, (43) is valid even for nonmonotone approximations and for more general operators of second order. In particular, it is valid for central difference approximations L_h of the convection-diffusion operators (6, 8), as long as h is small enough so that the real part of the eigenvalues of L_h is negative. This is due to the property of the θ -method for $0 \leq \theta < (1/2)$ that it is stable for operators with arbitrary spectrum in the left-half plane of the complex plane. Methods with such a property are called *A-stable* methods; see Dahlquist (1963). It is known that multistep time-discretization methods that are *A-stable* can be at most of second-order of approximation. On the other hand, certain implicit Runge-Kutta methods can have arbitrary high order and still be *A-stable*; see Axelsson (1964).

Remark 5. A simple method to estimate the discretization errors numerically is provided by Richardson extrapolation.

In applying this method to a nonstationary partial differential equation problem, we first keep k constant and run the problem over a time-step with a mesh in space with the following mesh sizes: say h and $(1/2)h$ (independently of each other). Then, as usual, $(1/3)[4v_{h/2}(x_i, t) - v_h(x_i, t)]$ provides an estimate of the space-discretization error, provided that this is $O(h^2)$, $h \rightarrow 0$.

Similarly, we can let h be fixed and run the numerical method with stepsize k and two steps with $k/2$. In the same manner as above, we can then estimate the time-discretization error.

If the solution is smooth, these estimates are frequently accurate. However, in an initial transient phase, for instance, when the solution is less smooth, the estimates are less accurate.

4.4 Nonlinear parabolic problems

We consider nonlinear evolution equations of parabolic type, that is, for which an asymptotic stability property holds.

The stability and discretization error for infinite time for the θ -method can be analyzed for nonlinear problems

$$\frac{du}{dt} + F(t, u) = 0, \quad t > 0, \quad u(0) = u_0 \in V \quad (44)$$

where V is a reflexive Banach space and $F: V \rightarrow V'$, where V' denotes the space dual to V . We have $V \hookrightarrow H \hookrightarrow V'$ for some Hilbert space H , where \hookrightarrow denotes continuous injections. Let (\cdot, \cdot) be the scalar product in H and $\|\cdot\|$ the associated norm. We assume that $F(\cdot, t)$ is *strongly monotone* (or *dissipative*), that is,

$$(F(t, u) - F(t, v), u - v) \geq \rho(t)\|u - v\|^2 \quad (45)$$

for all $u, v \in V$, $t > 0$, where $\rho \geq \rho_0 > 0$. It follows then

$$\frac{1}{2} \frac{d}{dt} (\|u - v\|^2) = -(F(t, u) - F(t, v), u - v) \leq -\rho_0 \|u - v\|^2, \quad t > 0$$

or

$$\|u(t) - v(t)\| \leq \exp(-\rho_0 t) \|u(0) - v(0)\|$$

For such nonlinear problems of parabolic type, the stability and discretization error for the θ -method can be analyzed (cf. Axelsson, 1984). Here, only a one-sided Lipschitz constant enters in the analysis. The accuracy of the θ -method, however, is limited to second order, at most. Here, we consider an alternative technique, which allows

an arbitrary high order of approximation. It is based on a boundary value technique.

The method is presented in a finite dimensional space $V = \mathbb{R}^n$ and on a bounded time interval. Given a system of differential equations

$$\frac{du}{dt} + \tilde{F}(t, u) = 0, \quad 0 < t \leq T, \quad u(0) = u_0 \text{ prescribed}$$

we first make a transformation of the equations to a more suitable form. In many problems, there may exist positive parameters $\varepsilon_i, 0 < \varepsilon_i \leq 1$, such that parts of \tilde{F} and of the corresponding parts of the Jacobian matrix $\partial \tilde{F} / \partial u$ are unbounded as $O(\varepsilon_i^{-1}), \varepsilon_i \rightarrow 0$. We then multiply the corresponding equation by this parameter to get

$$\varepsilon \frac{du}{dt} + F(t, u) = 0, \quad 0 < t \leq T \quad (46)$$

where ε is a diagonal matrix with entries ε_i , and now it is assumed that F and $(\partial F / \partial u)$ are bounded with respect to ε .

From the monotonicity of F for $t \geq t_0 > 0$ follows

$$\frac{1}{2} \frac{d}{dt} (\varepsilon(u-v), u-v) = - (F(t, u) - F(t, v), u-v) \leq -\rho(t) \|u-v\|^2 \leq -\rho(t) (\varepsilon(u-v), u-v), \quad t \geq t_0$$

so

$$\|u(t) - v(t)\|_t^2 \leq \exp \left(\int_{t_0}^t -2\rho(s) ds \right) \|u(t_0) - v(t_0)\|_{t_0}^2 \leq \|u(t_0) - v(t_0)\|_{t_0}^2, \quad t_0 \leq t \leq T$$

where $\|v\|_t = (\varepsilon v, v)^{1/2}$. This means that the system is contractive for $t \geq t_0$. In the initial phase $t \in (0, t_0)$, the system does not have to be contractive, that is, the eigenvalues of the Jacobian may have positive real parts there. In this interval, we may choose a step-by-step method with sufficiently small step sizes, in order to preserve stability and to follow the transients of the solution.

4.4.1 Boundary value techniques for initial value problems

The traditional numerical integration methods to solve initial value problems

$$\frac{du}{dt} = f(t, u(t)), \quad t > 0, \quad u(0) = u_0 \text{ given} \quad (47)$$

are step-by-step methods. Some such familiar methods are Runge-Kutta and linear multistep methods. In step-by-step methods, the error at the current time-step depends on the

local error at this step and also on the errors at all previous time-steps. In this way, the errors accumulate and the total error is typically larger by a factor $O(k^{-1})$ than the local errors. A direct global error control cannot be justified since the global error depends on the stability of the problem and the errors at the previous time-points.

In boundary value methods, on the other hand, all approximations at all time-points are computed simultaneously and such methods may be coined *global integration methods*. By its very nature, a boundary value method is better adapted for global error control.

For problems where one envisions that the solution suddenly may become unsmooth, one can implement boundary value methods as time-stepping methods but with much larger time-steps than for standard step-by-step methods.

A boundary value method can be composed of a sequence of forward step-by-step methods followed by a stabilizing backward-step method. As a simple example, let u_0 be given and

$$u^{n+1} - u^{n-1} - 2kf(t_n, u^n) = 0, \quad n = 1, 2, \dots, N-1 \\ u^N - u^{N-1} - kf(t_N, u^N) = 0$$

whose solution components u^1, u^2, \dots, u^N must be computed simultaneously. Such a method was analyzed in Axelsson and Verwer (1984). For a more recent presentation of boundary value and other methods, see Brugnano and Trigiante (1998).

The *Galerkin method*. For the interval (t_0, T) , the following global Galerkin method can be used. The interval is divided into a number of subintervals $(t_{i-1}, t_i), i = 1, 2, \dots, N$, where $t_N = T$. The length of the intervals, $t_i - t_{i-1}$, may vary smoothly with some function $h(t_i)$, but for ease of presentation, we assume that the intervals have the same length $t_i - t_{i-1} = h, i = 1, 2, \dots, N$. Consider each interval as an element on which we place some extra nodal points, $t_{i,j}, j = 0, 1, \dots, p$, such that $t_{i,j} = t_i + \xi_j h$, where ξ_j are the Lobatto quadrature points satisfying $0 = \xi_0 < \xi_1 < \dots < \xi_p = 1$ and $\xi_j + \xi_{p-j} = 1$. Hence, the endpoints of the interval are always nodal points and (if $p > 1$) we also choose $p-1$ disjoint nodal points in the interior of each element.

To each nodal point, we associate a basis function $\phi_{i,j}$. The basis functions may be exponential or trigonometric functions, and may also be discontinuous, but here we consider the case that they are continuous and polynomial over each element. Basis functions corresponding to interior nodes have support only in the element to which they belong, and those corresponding to endpoints have a support over two adjacent elements (except those at t_0 and t_N). The number of nodal points in each closed interval then equals the degree of the polynomial p plus one.

Let S_h^0 be the subspace of test functions that are zero at t_0 , that is,

$$S_h^0 = \text{span} \{ \phi_{i,j}, i = 0, 1, 2, \dots, N-1, j = 1, 2, \dots, p \}$$

Let

$$a(U; V) = \int_{t_0}^T \left(\varepsilon \frac{dU}{dt} + F(t, U), V \right) dt, \\ U, V \in [H^1(t_0, T)]^m$$

where $H^1(t_0, T)$ is the first-order Sobolev space of functions with square-integrable derivatives. To get an approximation \tilde{U} of the solution of (46), we take a vectorial test function $V = \phi_{i,j}^{[r]}$, multiply the equation, and after integration, we obtain

$$a(\tilde{U}; \phi_{i,j}^{[r]}) = \int_{t_{i-1}}^{t_i} \left(\varepsilon \frac{d\tilde{U}}{dt} + F(t, \tilde{U}), \phi_{i,j}^{[r]} \right) dt = 0, \\ i = 1, \dots, N-1 \\ j = 0 \quad (48)$$

$$a(\tilde{U}; \phi_{i,j}^{[r]}) = \int_{t_i}^{t_{i+1}} \left(\varepsilon \frac{d\tilde{U}}{dt} + F(t, \tilde{U}), \phi_{i,j}^{[r]} \right) dt = 0, \\ i = 0, 1, \dots, N-1 \\ j = 1, 2, \dots, p-1 \\ \text{and at } t_N = T, \text{ we get} \quad (49)$$

$$a(\tilde{U}; \phi_{N,0}^{[r]}) = \int_{t_{N-1}}^{t_N} \left(\varepsilon \frac{d\tilde{U}}{dt} + F(t, \tilde{U}), \phi_{N,0}^{[r]} \right) dt = 0 \quad (50)$$

We choose in turn $\phi_{i,j}^{[r]} = \phi_{i,j} e_r$, where e_r is the r th coordinate vector. This defines the Galerkin approximation \tilde{U} corresponding to S_h^0 , where

$$\tilde{U} = U(t_0)\phi_{0,0} + \sum_{i=0}^{N-1} \sum_{j=1}^p d_{i,j} \phi_{i,j}, \quad d_{i,j} \in \mathbb{R}^m$$

that is, we have imposed the essential boundary conditions at t_0 . Clearly,

$$a(U; V) = 0 \quad \forall V \in [H^1(t_0, T)]^m$$

From (48), we obtain

$$a(U; V) - a(\tilde{U}; V) = \int_{t_0}^{t_{i+1}} \left[\varepsilon \left(\frac{dU}{dt} - \frac{d\tilde{U}}{dt} \right) + (F(t, U) - F(t, \tilde{U})), V \right] dt = 0$$

$$V = \phi_{i,j}^{[r]}, \quad j = 0, i = 0, 1, \dots, N-1, \\ r = 1, 2, \dots, m \quad (51)$$

and similarly for (49) and (50).

To estimate the Galerkin discretization error $U - \tilde{U}$, we let $U_i \in S_h$ be the interpolant to U on $\{t_{i,j}\}, i = 0, 1, \dots, N-1$. From the representation $U - \tilde{U} = (U - U_i) + (U_i - \tilde{U}) = \eta - \theta$, we see that $\theta = U_i - \tilde{U} \in S_h^0$. Assuming that the solution U is sufficiently smooth, from the interpolation error expansion in integral form, we obtain the usual Sobolev norm estimates for the interpolation error:

$$\int_{t_0}^T \|U - U_i\|^2 dt \leq C_0 h^{2(p+1)} \int_{t_0}^T \|U\|_{p+1}^2 dt \\ \int_{t_0}^T \left\| \frac{dU}{dt} - \frac{dU_i}{dt} \right\|^2 dt \leq C_1 h^{2p} \int_{t_0}^T \left\| \frac{dU}{dt} \right\|_{p+1}^2 dt \quad (52)$$

Here, $\|U\|_{p+1}^2 = \int_{t_0}^T \sum_{i=0}^{p+1} |(\partial^i U / \partial t^i)|^2 dt$ is the norm in the Sobolev space $H^{p+1}(t_0, T)$.

Theorem 14. Let U be the solution of (46), and conditions

$$(F(t, U) - F(t, V), U - V) \geq \rho(t) \|U - V\|^2 \\ \|F(t, U) - F(t, V)\| \leq C \|U - V\|, \quad t > 0$$

be satisfied. Then the Galerkin solution \tilde{U} , in the space of piecewise polynomial continuous functions of degree p , defined by (48-50) satisfies

$$\|U - \tilde{U}\| = O(h^{p+1}) \left\{ \|U\|_{p+2}^2 + \|U\|_{p+1}^2 \right\}^{1/2}, \quad h \rightarrow 0$$

where $v = 1$ if $p = 1, 1 \geq v \geq (1/2)$ if $p = 3, 5, \dots$ and $v = 0$ if p is even, and

$$\|V\|^2 = \frac{1}{2} (\varepsilon V(T), V(T)) + \int_{t_0}^T \rho(t) \|V(t)\|^2 dt$$

Proof. For a detailed proof, see the supplement of (Axelsson and Verwer, 1984). \square

We have assumed that F is Lipschitz-continuous, that is, $\|F(t, u) - F(t, v)\| \leq C \|u - v\|$ for all $u, v \in \mathbb{R}^m$. This is, however, a severe limitation as it is a two-sided bound.

Difference schemes. In order to get a fully discretized scheme, one has to use numerical quadrature, which results in various difference schemes. We consider this only for $p = 1$. Then, $\phi_{i,p} = \phi_i$ are the usual hat functions and there are no interior nodes. With $\tilde{U} = U(t_0)\phi_0 + \sum_{i=1}^N U_i \phi_i$,

relations (48), (50) imply

$$\begin{cases} \varepsilon(\tilde{U}_{i+1} - \tilde{U}_{i-1}) = 2 \int_{t_{i-1}}^{t_{i+1}} F(t, \tilde{U}_{i-1}\phi_{i-1} + \tilde{U}_i\phi_i \\ \quad + \tilde{U}_{i+1}\phi_{i+1})\phi_i dt \\ \varepsilon(\tilde{U}_N - \tilde{U}_{N-1}) = \int_{t_{N-1}}^{t_N} F(t, \tilde{U}_{N-1}\phi_{N-1} + \tilde{U}_N\phi_N)\phi_N dt \end{cases} \quad (53)$$

We call this the *generalized midpoint rule difference scheme*. Let $F_i = F(t_i, \tilde{U}_i)|_{t=t_i}$. If we use numerical integration by the trapezoidal rule, that is,

$$\int_{t_{i-1}}^{t_i} F\phi_i dt \approx \frac{h}{2}[F_{i-1}\phi_i(t_{i-1}) + F_i\phi_i(t_i)] = \frac{1}{2}hF_i$$

which is known to be of $O(h^2)$ accuracy, on the basis of (53), we may derive a more accurate difference scheme. For this purpose, let

$$F(t) \approx \frac{1}{2}(F_{i-1} + F_i) + \left(t - t_i + \frac{h}{2}\right) \frac{1}{h}(F_i - F_{i-1}),$$

$$t_{i-1} \leq t \leq t_i$$

except that for the last formula in (53), we use $F(t) \approx (1/2)(F_{N-1} + F_N)$, $t_{N-1} \leq t \leq t_N$. Then,

$$\begin{aligned} \int_{t_{i-1}}^{t_i} F\phi_i dt &\approx \frac{h}{4}(F_{i-1} + F_i) + \frac{h}{12}(F_i - F_{i-1}) \\ &= \frac{h}{6}(F_{i-1} + 2F_i), \quad i = 1, 2, \dots, N-1 \end{aligned}$$

and similarly

$$\int_{t_i}^{t_{i+1}} F\phi_i dt \approx \frac{h}{6}(F_{i+1} + 2F_i)$$

Hence, the generalized midpoint rule (53) takes the form

$$\begin{cases} \varepsilon(\tilde{U}_{i+1} - \tilde{U}_{i-1}) = \frac{h}{3}(F_{i-1} + 4F_i + F_{i+1}), \\ \quad i = 1, 2, \dots, N-1, \\ \varepsilon(\tilde{U}_N - \tilde{U}_{N-1}) = \frac{h}{2}(F_{N-1} + F_N) \end{cases} \quad (54)$$

We notice that this is a combination of the Simpson and trapezoidal rules.

For this combination, numerical tests in Axelsson and Verwer (1984) indicate very accurate results. It is found that already on a very coarse mesh ($h = 1/4$), the accuracy is high. For $(du/dt) = \delta u$ and $\delta < 0$, the order of convergence seems to be ≈ 3.5 .

The above method is related to certain types of implicit Runge-Kutta methods; see, for example, Axelsson (1964) and Butcher (1977). As such, they are A-stable. The global coupling preserves this stability.

4.5 Computational aspects

The θ -method, $\theta \neq 1$ and other unconditionally stable methods are implicit, that is, they give rise to a linear system of algebraic equations to be solved at each time-step. For this purpose, one can use direct or iterative solution methods. Hereby, the associated cost is important, in particular as there will be a large number ($O(k^{-1})$) of such systems to be solved.

4.5.1 Iterative solution methods. The conjugate gradient method

The matrix in the linear system, which arises in the θ -method and has the form $I - (1 - \theta)kA$. If A is symmetric and negative definite, which is typical for parabolic problems, then the system has a condition number

$$\kappa = \frac{1 + (1 - \theta)kb}{1 + (1 - \theta)ka}$$

where $a = -\max_j \lambda_j$ and $b = -\min_j \lambda_j$ and λ_j are the eigenvalues of A . Hence, the condition number is bounded by $\kappa < 1 + (1 - \theta)kb$, that is (if $a \leq 1$), the condition number of A is typically multiplied by $(1 - \theta)k/a$, which can significantly decrease its value. For difference methods for second-order operators \mathcal{L} , it holds $b = O(h^{-2})$, so the number of iterations using the standard unpreconditioned conjugate gradient method grows only as $O(k^{1/2})$, if $k = O(h)$. Hence, the arising systems can normally be solved with an acceptable expense, in particular if one, in addition, uses a proper preconditioning of the matrix.

For stability reasons, for an explicit method, one must let $k = O(h^2)$. Hence, there are $O(h^{-1})$ more time-steps, so explicit methods are generally more costly except where there is a need to choose small time-steps of $O(h^2)$ due to approximation accuracy reasons. However, for $k = O(h^2)$, this is balanced by the condition number $\kappa = O(1)$, that is, the number of iterations are bounded irrespective of h , which make implicit methods still preferable because of their robust stability.

For problems with a small condition number, approximate inverse preconditioners can be quite accurate; see, for example, Axelsson (1994) for references to such methods. This can make implicit methods behave essentially as explicit but with no time-step stability condition. An example is the Euler backward method, where the arising matrix, $I - kA$, can be approximated by $(I - kA)^{-1} \approx I + kA$, the first term in the Neumann expansion. This is equivalent to the Euler forward method. Adding more terms increases the accuracy of this approximation if $\|kA\| < 1$.

Frequently, it can be efficient to reuse the same preconditioner for several time-steps, whence its construction cost can be amortized.

4.5.2 Periodic forcing functions

For problems

$$u_t = \Delta u + f(x, t), \quad x \in \Omega, \quad t > 0$$

where the forcing function is periodic in time, that is, $f(x, t) = f_0(x)e^{i\omega t}$, one can apply the Ansatz

$$u(x, t) = v(x, t)e^{i\omega t}$$

where u and v are complex-valued functions. We find then,

$$v_t = \Delta v - i\omega v = f_0(x)$$

Using the θ -method or any other implicit method, we must solve a system of the form

$$(A + i\omega I)(\xi + i\eta) = a + ib$$

where $A = I - k(1 - \theta)\Delta$, and ξ, η, a, b are real vectors. Multiplying by $A - i\omega I$, we get

$$(A^2 + \omega^2 I)(\xi + i\eta) = Aa + \omega b + i(Ab - \omega a)$$

or

$$\left[I + \left(\frac{1}{\omega} A \right)^2 \right] \xi = \frac{1}{\omega^2} Aa + \frac{1}{\omega} b \quad (55)$$

and a similar system for η . Equation (55) can be solved efficiently by iteration using a preconditioning in the form $[I + (1/\omega)A]^2$; see, for example, Axelsson and Kucherov (2000). Here, the condition number bound does not depend on $(1/\omega)A$.

4.5.3 Direct solution methods

Direct solution methods can be an efficient alternative to implicit methods if one uses constant stepsizes, and the coefficients in the differential operator do not depend on time, implying that the matrix arising at each time-step is constant. It can hence be factored, for instance, in triangular factors, once and for all, and the arising cost at each time-step comes from solving the factored systems only. For some problems, a nested dissection ordering scheme can be an efficient alternative to standard methods. For further details; see, for example, Axelsson and Barker (1984) and George and Liu (1981).

In general, however, the cost and demand of memory for direct solution methods grow more than linearly with increasing size of the system and they cannot be considered to be a viable choice for very large scale systems.

4.5.4 Alternating direction implicit methods

The alternating direction implicit method (ADI); see, for example, Peaceman and Rachford (1955) and the similar fractional step method (Yanenko, 1971) can sometimes be used to solve difference equations for elliptic problems more efficiently than some standard methods can. However, they can also be seen as special time-splitting methods using a direct solution method for the arising systems, which is the aspect of the ADI methods to be dealt with here.

For notational simplicity, consider a homogeneous equation

$$u_t + Au = 0, \quad t > 0, \quad u(0) = u_0 \quad (56)$$

where $A = A_1 + A_2$ is the sum of two positive definite operators and it is assumed that systems with $I + \alpha A_i$, $i = 1, 2$, $\alpha > 0$ can be solved more efficiently than systems with $I + \alpha A$. This holds, for example, if $A_1 u = -(\partial^2 u / \partial x^2)$ and $A_2 u = -(\partial^2 u / \partial y^2)$ since the one-dimensional problems $I + \alpha A_i$, after discretization and standard ordering, become tridiagonal matrices that can be solved with an optimal order of computational complexity (proportional to the order of the systems).

To derive the ADI methods, we utilize the Crank-Nicolson method for constant time-step k , which with $u_n = u(t_n)$, $t_n = nk$ takes the form

$$\left(I + \frac{k}{2} A \right) u_{n+1} = \left(I - \frac{k}{2} A \right) u_n, \quad n = 0, 1, \dots \quad (57)$$

Here, we add $(k/2)A_1 \cdot (k/2)A_2$ to both sides in (57) and rewrite it as

$$\begin{aligned} (I + \tilde{A}_1)(I + \tilde{A}_2)u_{n+1} &= (I - \tilde{A}_1)(I - \tilde{A}_2)u_n \\ &\quad + \tilde{A}_1 \tilde{A}_2 (u_{n+1} - u_n) \end{aligned} \quad (58)$$

where $\tilde{A}_i = (k/2)A_i$, $i = 1, 2$. Since for sufficiently smooth functions u , $u_{n+1} - u_n = O(k)$, the last term in (58) has the same order, $O(k^3)$, as the local truncation error of the Crank-Nicolson method. If we neglect this term, the resulting method becomes an alternating direction scheme of the form

$$\begin{aligned} (I + \tilde{A}_1)(I + \tilde{A}_2)u_{n+1} &= (I - \tilde{A}_1)(I - \tilde{A}_2)u_n, \\ n &= 0, 1, \dots \end{aligned} \quad (59)$$

where the two arising equations with matrices $I + \tilde{A}_1$ and $I + \tilde{A}_2$ can be solved with an optimal order computational complexity at each time-step.

For the stability analysis of it, we let $\tilde{u}_n = (I + A_2)u_n$ and rewrite it in the form

$$\tilde{u}_{n+1} = (I + \tilde{A}_1)^{-1}(I - \tilde{A}_1)(I - \tilde{A}_2)(I + \tilde{A}_2)^{-1}\tilde{u}_n$$

which is clearly a stable scheme, since $\|(I + \tilde{A}_1)^{-1}(I - \tilde{A}_1)\| < 1$, $i = 1, 2$.

Although the neglected term in (58) has the same order as the truncation error, it can still make an undesirably large perturbation on the discretization error. We therefore correct the time-stepping scheme by adding the term $\tilde{A}_1\tilde{A}_2(u_n - u_{n-1})$ to (59). If the solution is sufficiently smooth, then holds $\tilde{A}_1\tilde{A}_2(u_n - u_{n-1}) = \tilde{A}_1\tilde{A}_2(u_{n+1} - u_n) + O(k^4)$, so we have fully corrected for the last term in (58), which was missing in (59).

The resulting scheme is now a three-level method. To analyze its stability, we write it in the form

$$\begin{aligned} \tilde{u}_{n+1} &= (I + \tilde{A}_1)^{-1}(I - \tilde{A}_1)(I - \tilde{A}_2)(I + \tilde{A}_2)^{-1}\tilde{u}_n \\ &+ (I + \tilde{A}_1)^{-1}\tilde{A}_1\tilde{A}_2(I + \tilde{A}_2)^{-1}(\tilde{u}_n - \tilde{u}_{n-1}), \\ n &= 0, 1, \dots \end{aligned} \quad (60)$$

An elementary computation shows that

$$\begin{aligned} \tilde{u}_{n+1} &= [(I - 2\tilde{A}_1)(I - 2\tilde{A}_2) + \tilde{A}_1\tilde{A}_2]\tilde{u}_n \\ &- \tilde{A}_1\tilde{A}_2\tilde{u}_{n-1}, \quad n = 0, 1, \dots \end{aligned}$$

where $\tilde{A}_i = (I + \tilde{A}_i)^{-1}\tilde{A}_i$, $i = 1, 2$.

We assume now that A_1, A_2 commute (however, a quite restrictive assumption; see Varga, 1962). Then, the matrices $G = \tilde{A}_1\tilde{A}_2$ and $H = (I - 2\tilde{A}_1)(I - 2\tilde{A}_2) + \tilde{A}_1\tilde{A}_2$ are symmetric and, since \tilde{A}_i , $i = 1, 2$ are positive definite, their eigenvalues g, h are contained in the intervals $(0, 1)$ and $(-1, 2)$ respectively. An easy computation shows that the characteristic equation $\lambda^2 - h\lambda + g = 0$ has roots $|\lambda_i| < 1$, which also implies the stability of the modified scheme (60).

Remark 6. The given boundary conditions for the problem (56) are assumed to have been implemented in the matrix A and, through the splitting, in A_1 and A_2 also. The above ADI method is applicable for higher order, such as fourth order, space approximations as well, but in this case, the arising matrices $I + \alpha A_i$, $i = 1, 2$ are pentadiagonal or have even a higher order of structure. They can still be solved for with an optimal order of computational complexity. To get such a pentadiagonal structure for both the systems, an intermediate reordering of the unknowns must be applied for at least one of the systems.

5 FINITE DIFFERENCE METHODS FOR HYPERBOLIC PROBLEMS

In this section, we consider the numerical solution of scalar hyperbolic problems. There are two types of such equations of particular interest, namely,

(i) the first-order hyperbolic equation, typically of the form

$$au_x + bu_t = f, \quad 0 < x < 1, \quad t > 0$$

(ii) the second-order hyperbolic equation, typically of the form

$$u_{tt} = \Delta u + f, \quad 0 < x < 1, \quad t > 0$$

The latter, called the wave equation, describes transportation of waves, for instance, in acoustics, optics, and electromagnetic field theory (Maxwell's equation).

The numerical solution of the first-order problem can be seen as a special case of the more general convection-diffusion equation, and is discussed in Section 6.

In the present section, the wave equation is analyzed. We first prove that solutions of wave equations have quite a different behavior from solutions of parabolic equations, in the respect that the energy of a pure wave equation is constant (assuming no damping term and zero source function f), that is, the system is *conservative*. Furthermore, information (such as initial data) is transported with *finite velocity*. This is in contrast with a parabolic problem without sources, where the energy decays exponentially, and data at any point influences the solution at all interior points, that is, information is transported with infinite speed. We then discuss the stability of the method of lines for the numerical solution of wave equations and derive the familiar Courant-Friedrichs-Lewy (CFL) condition for the stability of the fully discretized equation.

Fourier expansion methods are limited to problems with constant coefficients. However, we choose in this section to illustrate their use on a fourth-order problem.

5.1 The wave equation

5.1.1 Derivation of the wave equation

To illustrate a problem that leads to a wave equation, consider an elastic string, which oscillates in a, say, vertical plane. Since the string is elastic, it has no stiffness for bending movements, so the stress S at a point acts along a line that is tangent to the string curve $u(x, t)$ at x for every t (see Figure 5). (Note that u is the transversal deflection, that is, the deflection in the direction (y), orthogonal to the

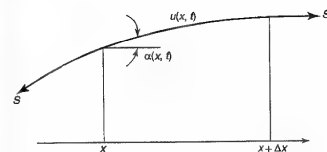


Figure 5. Stresses acting on a part of an elastic string.

direction x). We assume that the movement in each point is vertical, that is, horizontal movements are neglected. If the mass density per unit length is μ , then equating mass times acceleration with internal and external forces, and equating stresses in the transversal direction (Newton's law), we obtain

$$\begin{aligned} \mu \Delta x u_{tt} &= S(x + \Delta x, t) \sin \alpha(x + \Delta x, t) \\ &- S(x, t) \sin \alpha(x, t) + f \Delta x \end{aligned}$$

or, using $\sin \alpha(x, t) = [\tan \alpha / (\sqrt{1 + \tan^2 \alpha})] = [u_x / (\sqrt{1 + u_x^2})]$, we obtain the equation

$$\mu \Delta x u_{tt} = \left[S \frac{u_x}{\sqrt{1 + u_x^2}} \right]_{x-\Delta x}^{x+\Delta x} + f \Delta x$$

where f is an outer force, acting in the positive y -direction. Dividing by Δx and letting Δx go to zero, we obtain

$$\mu u_{tt} = \left(\frac{S u_x}{\sqrt{1 + u_x^2}} \right)_x + f, \quad t > 0 \quad (61)$$

This equation is nonlinear and has variable coefficients in general. If we assume that the amplitudes of the oscillations are small ($u_x^2 \ll 1$), and that μ and S are constant, then we get the linearized problem with constant coefficients,

$$\frac{1}{c^2} u_{tt} - u_{xx} = \frac{f}{S}, \quad t > 0 \quad (62)$$

where $c = (\sqrt{S/\mu})$.

If in addition, the Cauchy data, namely, the initial position $u(x, 0) = u_0(x)$ and the initial velocity $u_t(x, 0) = u_1(x)$ are known for all x ($-\infty < x < \infty$), we have a unique solution defined. In practice, of course, the string has a bounded length. If we assume that it is fixed at $x = 0$ and $x = 1$, say $u(0, t) = \alpha$, $u(1, t) = \beta$, then we have an example of a mixed initial boundary value problem.

Note that the second-order differential operator in (62) can be factorized into a product of two differential operators

of first order,

$$\frac{1}{c^2} \left(\frac{\partial}{\partial t} \right)^2 - \frac{\partial^2}{\partial x^2} = \left(\frac{1}{c} \frac{\partial}{\partial t} - \frac{\partial}{\partial x} \right) \left(\frac{1}{c} \frac{\partial}{\partial t} + \frac{\partial}{\partial x} \right)$$

This indicates that the solution consists of two waves transported with velocity c in the directions defined by the lines $x - ct = \text{constant}$ and $x + ct = \text{constant}$, that is,

$$u(x, t) = F(x - ct) + G(x + ct).$$

Here, F and G are defined by the initial data. For a mixed initial boundary value problem there may also appear reflected waves at the boundaries.

5.1.2 Domain of dependence

Consider the solution of the homogeneous problem $u_{tt} = c^2 u_{xx}$, $t > 0$, $-a < x < a$, where $c > 0$, $a \gg 1$, with boundary and initial conditions $u(-a, t) = u(a, t) = 0$, $t > 0$, $u(x, 0) = u_0(x)$, $u_t(x, 0) = u_1(x)$, for $-a < x < a$ and $t \leq (a - |x|)/c$.

If u_0 is twice continuously differentiable and u_1 is continuously differentiable, the explicit solution takes the form

$$u(x, t) = \frac{1}{2} [u_0(x - ct) + u_0(x + ct)] + \frac{1}{2c} \int_{x-ct}^{x+ct} u_1(s) ds \quad (63)$$

Hence, the solution $u(x_0, t_0)$ at some point (x_0, t_0) ($t_0 \leq (a - |x_0|)/c$) is only a function of the values of u_0 and u_1 on the interval $[x_0 - ct_0, x_0 + ct_0]$. For the inhomogeneous problem, $u_{tt} = c^2 u_{xx} + f$, one finds that the solution is the sum of the homogeneous solution (63) and a particular solution, which is equal to $(1/2)$ times the integral of f over the triangle in Figure 6. This is called the *domain of dependence* for the point (x_0, t_0) , and is defined as the smallest set $D(x_0, t_0)$ for points in the (x, t) plane such that $u_0, u_1, f = 0$ in $D(x_0, t_0)$ implies $u(x_0, t_0) = 0$. It is hence

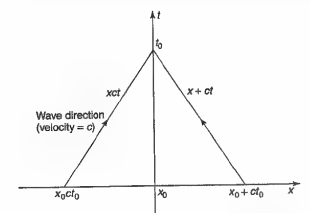


Figure 6. Domain of dependence for the wave equation.

bounded by the two characteristic lines through (x_0, t_0) and the x -axis.

5.1.3 Conservation of energy

We demonstrate the principle of conservation of energy for the string problem, where $u(-a, t) = u(a, t) = 0$ and $a \gg 1$. We assume then that u_0 is a smooth function with compact support and that the length of its support interval is much smaller than a . The solution consists then of two smooth waves with initial shape of $(1/2)u_0$, traveling with velocity c in the negative and positive direction, respectively (see Figure 7).

The energy has two sources,

- the kinetic energy $= (1/2) \int_{-a}^a \mu u_t^2 dx$, and the potential energy, which later is due to the stretching of the elastic string, that is,
- the potential energy $= \int_{-a}^a S \left(\sqrt{1 + u_x^2} - 1 \right) dx$.

Note that u_t is the velocity in the vertical direction. The total energy at time t is then

$$E(t) = \int_{-a}^a \left[\frac{1}{2} \mu u_t^2 + S \left(\sqrt{1 + u_x^2} - 1 \right) \right] dx \quad (64)$$

Theorem 15. For the solution of the homogeneous problem (61) (i.e., with $f \equiv 0$) with μ and S constant, the total energy is conservative, that is, E is constant. (In particular, the principle of minimal energy as for an elliptic problem is not valid!)

Proof. We have

$$E'(t) = \int_{-a}^a \left[\mu u_t u_{tt} + S \frac{d}{dt} \left(\sqrt{1 + u_x^2} \right) \right] dx \quad (65)$$

By (61) (with $f \equiv 0$), the first term equals

$$\begin{aligned} \int_{-a}^a \mu u_t u_{tt} dx &= \int_{-a}^a u_t \left(S \frac{u_x}{\sqrt{1 + u_x^2}} \right)_x dx \\ &= - \int_{-a}^a S u_{tx} \frac{u_x}{\sqrt{1 + u_x^2}} dx + \left[\frac{S u_t u_x}{\sqrt{1 + u_x^2}} \right]_{-a}^a \end{aligned} \quad (66)$$

where we have used partial integration.

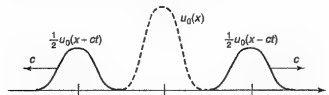


Figure 7. Two smooth traveling waves $((1/2)u_0(x-ct)$ and $(1/2)u_0(x+ct))$ in the linearized model.

The last term vanishes here because of the compact support of u_0 . (For simplicity, we consider only the case that t is small enough, so no reflected waves at the 'walls' at $-a$ and a appear.) Hence, by differentiation, it follows from (66) that

$$\int_{-a}^a \mu u_t u_{tt} dx = - \int_{-a}^a S \frac{d}{dx} \left(\sqrt{1 + u_x^2} \right) dx$$

so, by (65), $E'(t) = 0$. \square

When $u_x^2 \ll 1$, (61) takes the form $u_{tt} - c^2 u_{xx} = (c^2/S)f$, $t > 0$. Consider now the linearized wave equation in a bounded domain Ω and in a higher-space dimension (where the factor c^2/S is included in f),

$$u_{tt} = c^2 \Delta u + f(x, t), \quad t > 0, \quad x \in \Omega \in \mathbb{R}^d \quad (67)$$

with boundary conditions $u(x, t) = 0$, $x \in \partial\Omega$, $t > 0$ and initial conditions $u(x, 0) = u_0(x)$, $u_t(x, 0) = u_1(x)$, $x \in \Omega$. This equation is the linearized model for a vibrating elastic membrane, that is clamped along the boundary $\partial\Omega$ with an initial displacement, defined by u_0 and velocity u_1 (see Chapter 5, Volume 2).

5.1.4 Uniqueness of solutions

To prove uniqueness of solution of (67), we note that this is equivalent to proving that the homogeneous problem

$$u_{tt} = c^2 \Delta u \quad (68)$$

with homogeneous boundary and zero initial data has only the trivial solution. However, Theorem 15 implies that the energy $E(t)$ in (64) is constant, and because of the zero initial data $E(0) = 0$, that is, $E(t) = 0$. This shows that $u_t \equiv 0$ and $u_x \equiv 0$, so $u \equiv 0$.

Remark 7. Although the energy and velocity are constant, the shape of the two traveling waves in Figure 7 may change. Only for the linearized model is the shape unchanged. For the nonlinear (correct) model, the wave fronts tend to sharpen.

5.2 Difference approximation of the wave equation

5.2.1 A second-order scheme

For the time discretization of (67), we use a combination of the central difference approximation and a weighted average,

$$\begin{aligned} u(x, t+k) - 2u(x, t) + u(x, t-k) &= k^2 [\gamma u_{xx}(x, t+k) \\ &+ (1-2\gamma)u_{xx}(x, t) + \gamma u_{xx}(x, t-k)] \end{aligned} \quad (69)$$

$t = k, 2k, \dots$, where γ is a method parameter. In practice, we choose $0 \leq \gamma \leq (1/2)$. For the approximation of the Laplacian operator (Δ) , we use a difference approximation $(\Delta \approx \Delta_h)$, as described in Section 3. Then, the fully discretized difference scheme takes the form

$$\begin{aligned} [I - \gamma c^2 k^2 \Delta_h] [u_h(x, t+k) - 2u_h(x, t) + u_h(x, t-k)] \\ = c^2 k^2 \Delta_h u_h(x, t) + k^2 [\gamma f(x, t+k) \\ + (1-2\gamma)f(x, t) + \gamma f(x, t-k)], \quad t = k, 2k, \dots \end{aligned} \quad (70)$$

where u_h is the corresponding (differences) approximation of u .

Remark 8. If we use the method of lines, that is, replace Δ in (67) by Δ_h , then the resulting system is an initial value problem for a second-order, linear, ordinary differential equation, which can be rewritten as a system of first-order equations and subsequently discretized using the Crank-Nicolson method, certain types of implicit Runge-Kutta methods or multistep methods. Some details of analysis of such methods are presented in Section 4; see, Dahlquist (1978) and Hairer (1979) for further analysis. Naturally, after using the weighted difference approximation (69), we get the same difference scheme (70) as before.

Formula (70) describes a step-by-step method, which works on three time levels, $t+k$, t , $t-k$, at each step (a two-step method). Since (70) is a three-level scheme, we need starting values on the first two levels, $t=0$ and $t=k$. For $t=0$, we let $u_h(x, 0) = u_0(x)$. For $u_h(x, k)$, we may let

$$u_h(x, k) \approx u(x, 0) + ku_t(x, 0) = u_0(x) + ku_1(x)$$

or the higher-order approximation,

$$\begin{aligned} u_h(x, k) &\approx u(x, 0) + ku_t(x, 0) + \frac{1}{2}k^2 u_{tt}(x, 0) \\ &= u_0(x) + ku_1(x, 0) + \frac{1}{2}k^2 [c^2 \Delta u_0(x) + f(x, 0)] \end{aligned}$$

where we have used the differential equation and where we assume that u_0 is twice continuously differentiable.

For $\gamma \neq 0$, (70) is an implicit difference scheme, and at every time level, we must solve a linear system with matrix $I + \gamma c^2 k^2 \Delta_h$, where Δ_h is the difference matrix corresponding to $(-\Delta_h)$.

The computation of $u_h(x, t+k)$ is performed in two steps:

- Solve $\xi_h(x, t+k)$ from

$$(I + \gamma c^2 k^2 \Delta_h) \xi_h(x, t+k) = \text{r.h.s. of (70)}$$

- Compute

$$u_h(x, t+k) = 2u_h(x, t) - u_h(x, t-k) + \xi_h(x, t+k)$$

For $\gamma = 0$, on the other hand, we get the explicit scheme

$$\begin{aligned} u_h(x, t+k) &= 2u_h(x, t) - u_h(x, t-k) + c^2 k^2 \Delta_h \\ &\times u_h(x, t) + k^2 f(x, t), \quad t = k, 2k, \dots \end{aligned} \quad (71)$$

5.2.2 Stability analysis in L_2 -norm

Next, we analyze the stability of the completely discretized scheme (70). Clearly, we let u_h satisfy the given boundary conditions, that is, $u_h(x) = 0$, $x \in \partial\Omega_h$. Being a step-by-step method, perturbations of say initial data (70) could be unboundedly amplified as $t \rightarrow \infty$. To analyze this, we consider the homogeneous part of (70) (i.e. with $f \equiv 0$). Let λ_i , $v_i(x)$, $x \in \Omega_h$ be the eigensolutions of $(-\Delta_h)$, that is,

$$-\Delta_h v_i(x) = \lambda_i v_i(x), \quad x \in \Omega_h, \quad i = 1, 2, \dots, N$$

where $v_i(x) = 0$ on $\partial\Omega_h$. As we know from Section 3, $\lambda_i > 0$. (If $\Omega = [0, 1]^2$, then, in fact, $\lambda_{l,m} = (1/h^2)2(2 - \cos l\pi h - \cos m\pi h)$, $l, m = 1, 2, \dots, n$, where Ω_h is a $(n+2) \times (n+2)$ mesh.) To find a solution of (70) (with $f \equiv 0$), we consider the 'Ansatz',

$$u_h(x, t) = (\mu_i)^{t/k} v_i(x), \quad t = 0, k, 2k, \dots \quad (72)$$

where we assume that $u_0(x) = v_i(x)$, that is, it is an eigenfunction for some i , $1 \leq i \leq N$. Then, by substituting this Ansatz in (70) (with $f \equiv 0$), we get

$$\begin{aligned} (1 + \gamma c^2 k^2 \lambda_i)(\mu_i^2 - 2\mu_i + 1) &= \mu_i^{t/k-1} v_i(x) \\ &= -c^2 k^2 \lambda_i \mu_i^{t/k} v_i(x), \quad t = k, 2k, \dots \end{aligned}$$

Hence,

$$\mu_i^2 - 2\mu_i + 1 = -\mu_i \tau_i \quad (73)$$

where

$$\tau_i = \frac{c^2 k^2 \lambda_i}{1 + \gamma c^2 k^2 \lambda_i} \quad (74)$$

so

$$(\mu_i)_{1,2} = 1 - \frac{1}{2}\tau_i \pm \sqrt{\left(\frac{1}{2}\tau_i\right)^2 - \tau_i}$$

Theorem 16. The homogeneous difference scheme (70) (with $f \equiv 0$) is stable if $\tau_i = c^2 k^2 \lambda_i / (1 + \gamma c^2 k^2 \lambda_i) \leq 4 +$

$O(k^2)$. Perturbations of initial data are not amplified if $\tau_i < 4$, that is, if $\gamma \geq (1/4)$, or if

$$(ck)^2 < \frac{4}{1-4\gamma} \max_i \lambda_i \quad (75)$$

where $\gamma < (1/4)$ and they are boundedly amplified, uniformly in $t = k, 2k, \dots$, if $\tau_i \leq 4 + O(k^2)$.

Proof. By definition, the linear difference scheme (70) (with $f = 0$) is stable if and only if the solution corresponding to any perturbation of the initial values is bounded for all $t > 0$. Since any perturbation of the initial values can be written as a linear combination of the eigenfunctions $v_i(x)$, $x \in \Omega_h$ (which form the complete space), it suffices to consider an arbitrary perturbation $u_0(x) = v_i(x)$ and $u_k(x, k) = \mu_i(t/k)u_0(x)$, $i = 1, 2, \dots, N$. From (72) and (73), it then follows that the corresponding solutions are bounded if and only if $|\mu_i| \leq 1 + \tau_i k$, for some $\tau_i > 0$ (the von Neumann stability condition). This is so, because then $|\mu_i|^{1/k} \leq e^{\tau_i} \forall t > 0$.

The product of the two roots of (73) is equal to 1. Hence, we see that $|\mu_i| = 1$ (i.e. the perturbations are not amplified) if and only if $((1/2)\tau_i)^2 \leq \tau_i$, that is, if and only if $\tau_i \leq 4$. By (74), this means that $(1/4)ck\lambda_i \leq 1 + \gamma c^2 k^2 \leq \lambda_i$, that is, either $\gamma \geq (1/4)$ or $(ck)^2 \lambda_i \leq [4/(1-4\gamma)]$. It is valid for any perturbation, that is, for any i , if and only if (75) holds. Further, it follows that $|\mu_i| \leq 1 + \tau_i k$, if and only if $[(1/2)\tau_i]^2 - \tau_i \leq (ck)^2[1 + O(1)]$, $k \rightarrow 0$. \square

Corollary 4. If $\gamma < (1/4)$, then we have to choose k small enough, as follows by (75). In particular, if $\gamma = 0$ (the explicit method (71)), then $ck \leq 2/\max_i \lambda_i^{1/2}$. Note that $\max_i \lambda_i = O(h^{-2})$. (For the unit square problem, $\max_i \lambda_i \leq 8h^{-2}$.) Hence, $k \leq (2/c)O(h)$, ($k \leq h/(\sqrt{2}c)$ for the unit square.) This is a much more reasonable condition than for the explicit (Euler) method for the heat equation, where the stability conditions was $k \leq O(h^2)$ ($k \leq (1/4)h^2$ for the unit square). In fact, as shown later in this section, to balance the time discretization and space discretization errors, we shall normally choose $k = O(h)$ anyway.

If $\gamma \geq (1/4)$, then the method (70) is unconditionally stable, that is, stable for any choice of k .

5.2.3 Discretization error estimates

To find the order of the discretization errors $u - u_h$ as $h(k) \rightarrow 0$, we begin by considering the truncation error. (This is done here for a two-dimensional problem. For other space dimensions, one gets corresponding results.)

From (70), after division by k^2 , we obtain the difference equation

$$L_{h,k}u_h := [I - \gamma c^2 k^2 \Delta_h] \times \frac{u_h(x, t+k) - 2u_h(x, t) + u_h(x, t-k)}{k^2 - c^2 \Delta_h u_h(x, t)} = \bar{f}(x, t)$$

where $\bar{f}(x, t) = \gamma f(x, t+k) + (1-2\gamma)f(x, t) + \gamma f(x, t-k)$.

Definition 4. The truncation error for (70) is $L_{h,k}u - \bar{f}(x, t) = L_{h,k}(u - u_h)$.

Applying Taylor expansions for the central difference approximations, and assuming that data and the solution are sufficiently smooth, the truncation error takes the following form,

$$\begin{aligned} L_{h,k}(u - u_h) &= [I - \gamma c^2 k^2 \Delta_h] \\ &\times \frac{u(x, t+k) - 2u(x, t) + u(x, t-k)}{k^2 - c^2 \Delta_h u(x, t)} \\ &- [\gamma f(x, t+k) + (1-2\gamma)f(x, t) + \gamma f(x, t-k)] \\ &= [I - \gamma c^2 k^2 \Delta_h] \left[u_{tt}(x, t) + \frac{k^2}{12} u_{ttt}(x, t) + O(k^4) \right] \\ &- c^2 \Delta_h u(x, t) - f(x, t) - \gamma k^2 f_{tt}(x, t) + O(k^4) \\ &= u_{tt}(x, t) - c^2 \Delta_h u(x, t) - f(x, t) \\ &+ \frac{k^2}{12} u_{ttt}(x, t) - \gamma c^2 k^2 (\Delta_h u)_{tt} - \gamma k^2 f_{tt}(x, t) \\ &- \frac{c^2 h^2}{12} (u_{tt}^{(4)} + u_{tt}^{(4)}) + O(k^4) + O(h^2 k^2) + O(h^4) \end{aligned}$$

By use of the differential equation $u_{tt} = c^2 \Delta_h u + f$, we get $u_{tt}^{(4)} = c^2 \Delta_h u_{tt} + f_{tt}$ and, therefore,

$$L_{h,k}(u - u_h) = \left(\frac{1}{12} - \gamma \right) k^2 u_{ttt}^{(4)}(x, t) - c^2 \frac{h^2}{12} \times (u_{tt}^{(4)} + u_{tt}^{(4)}) + O(k^4) + O(h^2 k^2) + O(h^4) \quad (76)$$

Theorem 17. Assume that u is sufficiently smooth and that the difference scheme (70) is stable, that is, $\gamma \geq (1/4)$ or $(ck)^2 \leq [4/(1-4\gamma)](\max_i \lambda_i)^{-1}$. Then, the discretization error of (70) satisfies

- (a) $\|u - u_h\| \leq CT(k^2 + h^2)$, $h, k \rightarrow 0$, $0 \leq t \leq T$
 (b) $\|u - u_h\| \leq CT(k^4 + h^2)$, $h, k \rightarrow 0$, $0 \leq t \leq T$, if $\gamma = (1/12)$.

Proof. Let $e_h = u - u_h$ be the discretization error. Then, using estimates similar to those in Theorem 3, using the

property that $I - \gamma c^2 k^2 \Delta_h$ is of strongly positive type, one obtains

$$\|(u - u_h)(\cdot, t)\| \leq CT \|L_{h,k}(u - u_h)\|, \quad 0 \leq t \leq T$$

for some constant C (independent of h, k, T). Now, (76) readily implies the theorem. \square

Remark 9. If $\gamma \neq (1/12)$, we see that to balance the order of the errors in space and time, we have to choose $k = O(h)$, $h \rightarrow 0$. In the case $\gamma = (1/12)$; however, we may choose $k = O(h^{1/2})$ to balance the errors. Unfortunately, with such a choice the stability condition (75) is violated if $k(h)$ is small enough.

One can show that if we use a nine-point discretization of Δ (in \mathbb{R}^2), and if we use proper weighted averages of $f(\bar{x}, t+k)$, $f(\bar{x}, t)$ and $f(\bar{x}, t-k)$ at $\bar{x} = (x+h)$, x and $(x-h)$, then we get a scheme with error $O(k^4) + O(h^4)$ for $\gamma = (1/12)$. The corresponding matrix is less sparse, however, than for the five-point discretization.

5.2.4 Computational aspects

The linear systems of equations arising in (70) for $\gamma \neq 0$ and when $k = O(h)$ have condition number $O(1)$, $h \rightarrow 0$. Therefore, their inverses can be approximated accurately by a sparse matrix and correspondingly a preconditioned iterative solution method will work essentially as an explicit method.

5.2.5 The CFL condition

Consider the explicit difference scheme (71), or

$$D_t^+ D_t^- u_{h,k}(x, t) = c^2 D_x^+ D_x^- u_{h,k}(x, t)$$

for the numerical solution of $u_{tt} = c^2 u_{xx}$, $-\infty < x < \infty$, $u(x, 0) = u_0(x)$, $u_t(x, 0) = u_1(x)$, $-\infty < x < \infty$.

It can be represented by a stencil shown in Figure 8, where $\rho = ck/h$. The domain of dependence (a triangle) of the solution $u(x, t)$ at a point (x, t) is defined by the interval $[x-ct, x+ct]$ and (x, t) . As follows from (63), if u_0 is changed at the endpoints or u_1 is changed in some point of this segment, then, in general, $u(x, t)$ is changed. It is readily established that the solution of the difference equation also has a finite domain of dependence, that is, a smallest set $D_h(x, t)$ such that $u_{h,k}(x, t) = 0$ if $u_0 = 0$ in $D_h(x, t)$, and defined by the interval $[x-mh, x+mh]$, if $t = mk$.

Hence, for convergence, $(u - u_h)(x, t)$, $h \rightarrow 0$, we realize that this interval must contain the interval for the

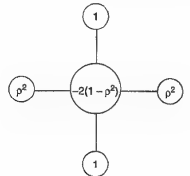


Figure 8. Difference stencil.

continuous domain of dependence, that is,

$$mh \geq ct = cmk = m\sqrt{\rho}h$$

Hence, we must have $\rho \leq 1$. (Note that this is in fact also sufficient for stability and hence also for convergence because the scheme is consistent – see (75) with $\gamma = 0$, where $\max_i \lambda_i = 4/h^2$ – for a problem in one-space dimension.) This condition is an example of a CFL condition: A necessary condition for the convergence of a difference approximation of a wave equation is that the discrete solution at any point has a domain of dependence that (at least for small values of k) covers the domain of dependence for the exact solution at the same point. Equivalently, we can state the CFL condition: The velocity (h/k) of the solutions of the discrete problem must be at least as big as the velocity c of the solution of the continuous problem.

Note that if an implicit method ($\gamma > 0$) is used, the domain of dependence of the difference method is the whole x -axis, because the inverse of the operator $I - \gamma c^2 k^2 \Delta_h$ is a full matrix. Hence, the above necessary condition is automatically satisfied.

5.2.6 Numerical dispersion

Related to the CFL condition is the numerical dispersion number d . We study this for the one-dimensional equation. If $d = (\omega/c)/\ell = 1$, the harmonic wave $u(x, t) = e^{i(\omega t - kx)}$ satisfies the homogeneous equation $u_{tt} = c^2 u_{xx}$. This relation between the frequency ω and the wave number must hold since all waves propagate with the same speed c .

For the explicit numerical scheme (70), with $\gamma = 0$, there holds

$$\begin{aligned} u_h(x, t+k) - 2u_h(x, t) + u_h(x, t-k) \\ = \rho^2 [u(x+h, t) - 2u(x, t) + u(x-h, t)] \end{aligned}$$

with $\rho = ck/h$. Let the wave number ℓ be fixed. The Ansatz $u_h(x, t) = e^{i(\omega_k t - \ell x)}$ shows the relation

$$e^{i\omega_k t} - 2 + e^{-i\omega_k t} = \rho^2 (e^{i\ell h/2} - 2 + e^{-i\ell h/2})$$

that is,

$$\left(e^{i\frac{\omega_k}{2}} - e^{-i\frac{\omega_k}{2}} \right)^2 = \rho^2 (e^{i\frac{\ell h}{2}} - e^{-i\frac{\ell h}{2}})^2$$

or

$$\sin \frac{\omega_k}{2} = \pm \rho \sin \frac{\ell h}{2} \quad (77)$$

It can be seen that $\omega_k = c\ell[1 - ((\ell h)^2/24)(1 - \rho^2) + O((\ell h)^4)]$. Hence, unless $\rho^2 = 1$, for the numerical solution there is a phase error $\omega - \omega_k$ and the angular frequency ω_k is only approximately proportional to the wave number. This means that the numerical solution of a wave package containing several different spatial frequencies will change shape as it propagates. The phenomenon of waves of differential frequencies traveling with different speed is called *dispersion*.

The number $d_h = (\omega_k/c|\ell|)$ shows how many grid cells the true solution propagates in one time-step. If $d_h = 1$, the spatial and temporal difference approximation cancel, and the signals propagate one cell per time-step, in either direction. If $d_h < 1$, on the other hand, the numerical dissipation can differ from the analytical by a significant amount, at least for the bigger wave numbers. If $d_h > 1$, the relation (77) yields complex angular frequencies for wave numbers such that $|\sin \ell h/2| > 1/|\rho|$. Therefore, some waves will be amplified exponentially in time, that is, the algorithm is unstable. This is in agreement with the violation of the stability conditions in Theorem 16. When $ck > h$, the signal of the true solution propagates more than one cell per time-step, which later is the propagation speed of the numerical solution of the explicit scheme. As we have seen, the CFL condition is then violated. Similar results hold for other explicit schemes.

5.3 A fourth-order problem

In this section, we analyze the following partial differential equation, which is a nonstationary model for the deflection of a thin plate, fixed at the boundary,

$$\frac{\partial^2 u}{\partial t^2} = -E \Delta^2 u + P, \quad t > 0, \quad (x, y) \in \Omega \subset \mathbb{R}^2$$

or

$$u = \frac{\partial u}{\partial t} = 0, \quad t > 0, \quad (x, y) \in \partial\Omega$$

$$u = \Delta u = 0, \quad t > 0, \quad (x, y) \in \partial\Omega \quad (78)$$

and with initial deflection and velocity of deflection,

$$u(x, y, 0) = u_0(x, y), \quad \frac{\partial u}{\partial t}(x, y, 0) = u_1(x, y), \quad (x, y) \in \Omega$$

where $u = u(x, y, t)$ is the deflection, $(\partial/\partial n)$ is the normal derivative, $P = P(x, y, t)$ is a pressure load, $E > 0$ is the stiffness coefficient, and Δ^2 is the biharmonic (fourth-order) operator,

$$\Delta^2 u = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^2 u = \frac{\partial^4 u}{\partial x^4} + 2 \frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} \quad (79)$$

The equation appears, for instance, in the modeling of the deflection (vibration) of spinning disks, such as 'floppy' disks, tapes, and thin beams; see, Benson and Bogy (1978) and Lamb and Southwell (1921) (see Chapter 5, Volume 2).

We first describe the behavior of solutions of (78) by use of a Fourier expansion. For the numerical solution, we use the method of lines, however, applied on a system of two equations, where each equation contains a Laplacian (second-order equation) instead of the biharmonic operator. By discretizations of the Laplacian operators we get a system of two coupled ordinary differential equations, which in their turn are discretized by the use of difference approximations. In the interest of brevity and clarity of exposition, we present the above analysis for a model problem in one space dimension. (Note also that we consider here the boundary condition of type $(\partial^2 u/\partial x^2) = 0$ instead of $(\partial u/\partial x) = 0$.)

Consider therefore

$$\frac{\partial^2 u}{\partial t^2} = -\frac{\partial^4 u}{\partial x^4} + P, \quad 0 < x < 1, \quad t > 0 \quad (80)$$

$$u(0, t) = u(1, t) = 0, \quad \frac{\partial^2 u}{\partial x^2}(0, t) = \frac{\partial^2 u}{\partial x^2}(1, t) = 0, \quad t > 0$$

$$u(x, 0) = g_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = g_1(x), \quad 0 < x < 1 \quad (81)$$

(where, for simplicity, we assume that the compatibility conditions $g_0(0) = g_0(1) = 0$, $g_1(0) = g_1(1) = 0$ hold and that $g_0 \in C^4(0, 1)$, $g_1 \in C^3(0, 1)$).

5.3.1 Fourier expansion

To find the behavior of the solutions of (80), we use the method of superposition and expansion of the solution in series of eigenfunctions (Fourier expansion). Consider first the homogeneous equation and the Ansatz $u(x, t) = \phi(t)\psi(x)$. We substitute into (80) (with $P = 0$) and obtain

$$\phi''(t)\psi(x) = -\phi(t)\psi^{(4)}(x)$$

or

$$\frac{\phi''(t)}{\phi(t)} = -\frac{\psi^{(4)}(x)}{\psi(x)} = -\lambda^2$$

which must be constant. We let $\lambda > 0$. (As shown below, the negative sign of $(-\lambda^2)$ turns out to be the correct choice.) Hence,

$$\phi(t) = \tilde{c}_1 e^{i\lambda t} + \tilde{c}_2 e^{-i\lambda t} = c_1 \cos \lambda t + c_2 \sin \lambda t$$

Similarly, we find that $\psi(x) = C_4 \cos(\sqrt{\lambda}x) + C_3 \sin(\sqrt{\lambda}x) + C_2 e^{\sqrt{\lambda}x} + C_1 e^{-\sqrt{\lambda}x}$.

From $u(0, t) = 0$ and $(\partial^2 u/\partial x^2)(0, t) = 0$, $(\partial^2 u/\partial x^2)(1, t) = 0$, we get $C_2 = C_1 = 0$, and from $u(0, t) = 0$, we get then $C_4 = 0$, and from $u(1, t) = 0$, $\psi(x) = C_3 \sin(\sqrt{\lambda}x)$. $(\sqrt{\lambda}) = (\sqrt{\lambda_k}) = k\pi$, $k = 1, 2, \dots$. Notice that this function also satisfies $\psi^{(2)}(0) = \psi^{(2)}(1) = 0$.

Hence, $u(x) = (c_1 \cos \lambda_k t + c_2 \sin \lambda_k t) \sin(k\pi x)$ satisfies the homogeneous equation and the homogeneous boundary conditions. By a proper choice of the Fourier coefficients $C_1^{(k)}, C_2^{(k)}$, we find that

$$u(x, t) = \sum_{k=1}^{\infty} (C_1^{(k)} \cos k^2 \pi^2 t + C_2^{(k)} \sin k^2 \pi^2 t) \sin(k\pi x)$$

also fulfills the two initial conditions, that is, u is a solution of (80) and (81) with $P = 0$. (If we expand P in such a Fourier series, we may also find a particular solution satisfying the inhomogeneous differential equation, but with homogeneous boundary and initial data.) Our conclusion is that the solution of (80), (81) consists of harmonic oscillations, both in time and space. The problem is of hyperbolic type.

5.3.2 The method of lines

Consider now the numerical solution of (80), (81) (for simplicity with $P = 0$). We rewrite the equation as a coupled system of two equations, each of second order only in space:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 v}{\partial x^2}$$

$$\frac{\partial v}{\partial t} = -\frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad t > 0$$

with initial values $u(x, 0) = g_0(x)$, $v(x, 0) = v_0(x)$, where $(\partial u/\partial t)(x, 0) = v_0' = g_1(x)$, $0 < x < 1$.

Note that we can compute v_0 from $v_0' = g_1(x)$ by integration.

The boundary conditions are $u(0, t) = u(1, t) = v(0, t) = v(1, t) = 0$, $t \geq 0$. The latter follows from $v_t =$

$-u_{xx} = 0$ for $x = 0$ and $x = 1 \forall t > 0$. Hence, v is constant for $x = 0$ and $x = 1$ for all $t > 0$, and we may put this constant to zero.

We approximate now $(\partial^2 v/\partial x^2)$ and $(\partial^2 u/\partial x^2)$ by central differences and we then get

$$\frac{d}{dt} \begin{bmatrix} u_h(x, t) \\ v_h(x, t) \end{bmatrix} = \begin{bmatrix} 0 & D^+ D^- \\ -D^+ D^- & 0 \end{bmatrix} \begin{bmatrix} u_h \\ v_h \end{bmatrix}, \quad t > 0, \quad x = h, 2h, \dots, 1-h \quad (82)$$

where $h = 1/(N+1)$. Let U_h, V_h be the vectors with components $u_h(x, t), v_h(x, t)$ respectively, at $x = x_k = kh$, $k = 1, \dots, N$. Then (82) may be rewritten in matrix form

$$\frac{d}{dt} \begin{bmatrix} U_h \\ V_h \end{bmatrix} = -h^{-2} \begin{bmatrix} 0 & B \\ -B & 0 \end{bmatrix} \begin{bmatrix} U_h \\ V_h \end{bmatrix}, \quad t > 0 \quad (83)$$

where

$$B = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ & \ddots & \ddots \\ 0 & -1 & 2 \end{bmatrix}$$

The eigenvalues of B are, as is well known,

$$\lambda_q = \left(2 \sin \frac{q\pi h}{2} \right)^2, \quad q = 1, 2, \dots, N$$

and the ones of $\begin{bmatrix} 0 & B \\ -B & 0 \end{bmatrix}$ are accordingly readily found to be

$$\mu_q = \pm i \lambda_q = \pm i \left(2 \sin \frac{q\pi h}{2} \right)^2, \quad i = (-1)^{1/2} \quad (84)$$

(Lagrange (1759), Thèse 'Propagation of sounds').

Hence, since μ_q is purely imaginary and the eigenvector space is complete, the solutions u_h, v_h of (82) are also bounded and are represented by wave (trigonometric) functions in time. Accordingly, the behavior of the method of lines describes the solution of (80) correctly, at least in a qualitative way.

5.3.3 Time discretization

To get a completely discretized equation, we approximate the time derivative in (82) by central differences, using the midpoint method. Hence, let

$$\frac{du_h(x, t)}{dt} \approx \frac{1}{2k} [u_h(x, t+k) - u_h(x, t-k)] \quad t = k, 2k, \dots$$

We get then

$$\begin{bmatrix} \tilde{U}_h(t+k) \\ \tilde{V}_h(t+k) \end{bmatrix} = \begin{bmatrix} \tilde{U}_h(t-k) \\ \tilde{V}_h(t-k) \end{bmatrix} - 2\frac{k}{h^2} \begin{bmatrix} 0 & B \\ -B & 0 \end{bmatrix} \times \begin{bmatrix} \tilde{U}_h(t) \\ \tilde{V}_h(t) \end{bmatrix}, \quad t = k, 2k, \dots \quad (85)$$

This is an explicit step-by-step method, working on three levels at a time (a two-step method). As initial values, we use

$$\begin{bmatrix} \tilde{U}_h(0) \\ \tilde{V}_h(0) \end{bmatrix} = \begin{bmatrix} g_0(x) \\ v_0(x) \end{bmatrix}, \quad \begin{bmatrix} \tilde{U}_h(k) \\ \tilde{V}_h(k) \end{bmatrix} = \begin{bmatrix} g_0(x) \\ v_0(x) \end{bmatrix} + k \begin{bmatrix} g_1(x) \\ -g_0'(x) \end{bmatrix} + \frac{k^2}{2} \begin{bmatrix} -g_0''(x) \\ -g_1'(x) \end{bmatrix}$$

Equation (85) is a homogeneous difference equation of second order. To analyze its stability, we make the Ansatz $r^m \mathbf{W}_q$, $m = 0, 1, \dots$ ($m = t/k$) for the solution, where \mathbf{W}_q is an eigenvector of $A = \begin{bmatrix} 0 & B \\ -B & 0 \end{bmatrix}$, that is, $A\mathbf{W}_q = \mu_q \mathbf{W}_q$. We then get $r^{m+1} \mathbf{W}_q = (r^{m-1} - 2\tau_q r^m) \mathbf{W}_q$, $m = 1, 2, \dots$ where

$$\tau_q = \frac{k}{h^2} \mu_q \quad (86)$$

Hence, r must solve the characteristic equation $r^2 = 1 - 2\tau_q$.

Let r_1, r_2 be its solutions. Note that $|r_1 r_2| = 1$ and $r_{1,2} = \tau_q \pm (1 - \tau_q^2)^{1/2}$. Since τ_q is purely imaginary, we have $\max_{m \geq 1} |r_1|^m \leq 1$ if and only if $|\tau_q| \leq 1$, that is, by (86) and (84), $k(2 \sin(q\pi h/2))^2 \leq h^2$. In fact, we then have $|r| = 1$. This must be satisfied for all $q = 1, 2, \dots, N$. Hence, $k \leq h^2/(2 \sin(N\pi/2(N+1)))^2$, or $k \leq (1/4)h^2$. (Courant, Friedrichs, Levy, 1928).

The latter condition on k is a severe restriction on the time-step length. To avoid this, one can use a proper, unconditionally stable form of an implicit method for (83). For instance, the θ -method is applicable, $0 \leq \theta \leq (1/2)$, and for $\theta = (1/2)$, we have a conservative scheme, that is, the corresponding eigenvalues μ_q are purely imaginary. To solve evolution equations with eigenvalues that can be purely imaginary, one needs, in general, A -stable methods, that is, methods that, when applied to the model problem $u_t = \lambda u$, are stable for all λ with $\operatorname{Re} \lambda \leq 0$.

6 CONVECTION-DIFFUSION PROBLEMS

6.1 The convection-diffusion equation

Consider the convection-diffusion problem

$$\mathcal{L}u = -\nabla \cdot (\varepsilon \nabla u) + \mathbf{v} \cdot \nabla u + cu = f, \quad \mathbf{x} \in \Omega, u = 0 \text{ on } \partial\Omega \quad (87)$$

where Ω is a bounded domain in \mathbb{R}^n with, for example, Dirichlet boundary condition $u = g$, $\mathbf{x} \in \partial\Omega$. We assume that $\varepsilon > 0$, $c \geq 0$ and that \mathbf{v} is a given velocity vector function, defined in Ω . Typically, \mathbf{v} carries (conveys) some relative concentration (mixing ratio) of a dilute chemically active solute in a neutral solvent (fluid), which moves in a closed vessel with the imposed mass flow field \mathbf{v} ; f is the chemical source term and c is the rate coefficient for removal in chemical reactions (see Chapter 24, this Volume and Chapter 7, Volume 3).

When $\|\mathbf{v}\| = \{v_1^2 + v_2^2\}^{1/2}$ is large (the problem is then said to have a large Reynolds number), the convective or hyperbolic part $\mathbf{v} \cdot \nabla u$ of the equation dominates and the solution follows essentially the characteristic lines. On the other hand, when $\|\mathbf{v}\|$ is not large, the diffusive part $-\Delta u$ dominates the behavior of the solution.

6.2 Finite differences for the convection-diffusion equation

Consider first a one-dimensional problem

$$-u_{xx} + v u_x = f, \quad 0 < x < 1, \quad u(0) = \alpha, \quad u(1) = \beta \quad (88)$$

Using central differences for both u_{xx} and u_x results in a difference operator L_h , which is not of positive type (see Section 3) if the so-called local Peclet number $Pe \equiv (vh/2) > 1$. Then, the discrete maximum principle does not hold for L_h .

When v is very large, satisfying the condition $Pe \leq 1$ would require a very fine mesh and, hence, a very large cost in solving the corresponding algebraic system. To get a positive-type scheme, one can use the so-called upwind (backward) differences for u_x , that is,

$$u_x(x_i) \approx \frac{u(x_i) - u(x_{i-1}))}{h}, \quad \text{if } v > 0$$

$$\text{and} \quad u_x(x_i) \approx \frac{u(x_{i+1}) - u(x_i)}{h}, \quad \text{if } v < 0 \quad (89)$$

However, the leading local truncation error term now becomes $-v(h/2)u_{xx}$, which can be seen to cause extra

(artificial) diffusion, and has an approximation error that is of one order less than for the central difference scheme. When v is constant, $f \equiv 0$ and $\alpha = 0, \beta = 1$, the solution of (88) equals

$$u(x) = \frac{e^{v(x-1)} - e^{-v}}{1 - e^{-v}}$$

Hence, when v is large, then $u(x) \approx 0$, except in a thin layer of width $O(1/v)$ near the boundary point, where the solution increases rapidly. The central difference scheme,

$$\frac{1}{h^2}(-\tilde{u}_{i-1} + 2\tilde{u}_i - \tilde{u}_{i+1}) + v \frac{1}{2h}(\tilde{u}_{i+1} - \tilde{u}_{i-1}) = 0$$

$$i = 1, 2, \dots, n$$

or

$$-\left(1 + \frac{vh}{2}\right)\tilde{u}_{i-1} + 2\tilde{u}_i - \left(1 - \frac{vh}{2}\right)\tilde{u}_{i+1} = 0$$

has a solution \tilde{u}_i , which, if $(vh/2) \neq 1$, satisfies

$$\tilde{u}_i = C_1 \lambda_1^i + C_2 \lambda_2^i$$

where $\lambda_{1,2}$ are the roots of the equation $-[1 + (vh/2)] + 2\lambda - [1 - (vh/2)]\lambda^2 = 0$. Hence, $\lambda_1 = 1$ and $\lambda_2 = [1 + (vh/2)]/[1 - (vh/2)]$. Using the given boundary values, one finds

$$\tilde{u}_i = \frac{(-1)^{n-i+1}(1-\delta)^{n-i+1} - (-1)^{i+1}(1-\delta)^{i+1}}{1 - (-1)^{n+1}(1-\delta)^{n+1}}$$

with $\delta = 2/[1 + (vh/2)]$. This exhibits an oscillatory behavior, $\tilde{u}_{n+1} = 1$, $\tilde{u}_n \approx -1 + \delta$, $\tilde{u}_{n-1} \approx (1-\delta)^2$, and so on. On the other hand, considering only each second point, $i = n+1, n-1, n-3, \dots$, the solution takes the form

$$\tilde{u}_i = (1-\delta)^{n-i+1} \frac{1 - (-1)^{i+1}(1-\delta)^i}{1 - (-1)^{n+1}(1-\delta)^{n+1}}$$

that is, exhibits no oscillations. Similarly, for the upwind scheme, the solution turns out to be

$$\tilde{u}_i = \frac{(1+vh)^i - 1}{(1+vh)^{n+1} - 1}, \quad 0 \leq i \leq n+1$$

which shows that the solution is smeared out, that is, the width of the numerical layer is much bigger than that for u , typically $O(1/vh)$ instead of $O(1/v)$ when $v \gg 1$.

Frequently, it can be efficient to use central differences everywhere, except in the layer region, where the solution can be resolved using an adapted mesh with $h < 2/v$.

Alternatively, as shown in Section 6.6, one can use a generalized difference scheme based on local Green's functions.

As for the one-dimensional problem, for higher-dimensional problems, we construct a mesh Ω_h in Ω and

consider two approximations for \mathcal{L} at every point $(x, y) \in \Omega_h$

- central differences (of Shortley-Weller-type next to a curved boundary) for $-\Delta u$, u_x and u_y . The corresponding operator is denoted by $L_h^{(0)}$.
- Central differences for $-\Delta u$ but "upwind" differences for u_x and u_y , that is,

$$\frac{\partial u}{\partial x}(x, y) \approx D_x^- u = \frac{u(x, y) - u(x-h, y)}{h}, \quad \text{if } v_1(x, y) > 0$$

$$\frac{\partial u}{\partial x}(x, y) \approx D_x^+ u = \frac{u(x+h, y) - u(x, y)}{h}, \quad \text{if } v_1(x, y) < 0$$

and similarly for u_y . The corresponding difference operator is denoted by $L_h^{(1)}$.

The first discretization method gives local truncation errors $O(h^4)$ (except possibly near a curved boundary), while the second difference method has local errors $O(h^3)$ at best, and one can expect that the first method gives more accurate solutions for sufficiently small values of h . However, when $\|\mathbf{v}\|$ is large, the first scheme is not of positive type and the difference matrix may not be monotone. This causes numerical solutions with oscillatory behavior when the solution of (87) has steep gradients in parts of the domain Ω . On the other hand, the second scheme always gives a difference approximation of positive type and such an operator can not cause such nonphysical wiggles. However, the first-order approximation of u_x and u_y causes truncation error terms

$$-v_1 \frac{h^3}{2} u_{xx} - v_2 \frac{h^3}{2} u_{yy}$$

which, as can be seen, corresponds to added diffusion terms. This means that the numerical solution is smeared out, where there occurs sharp gradients in the solution of (87). Hence, both schemes have advantages and disadvantages. As it turns out, the first scheme works well when we use (locally) a sufficiently fine difference mesh where steep gradients occur but not necessarily in other parts of the domain.

A non-self-adjoint problem

$$\mathcal{L}u = -\sum_{i=1}^n a_i u_{x_i} + \sum_{i=1}^n v_i u_{x_i} + cu = f \quad (90)$$

where a_i and v_i are constants and $a_i > 0$, can be transformed into a self-adjoint form. Namely, with

$u = w \exp(1/2) \sum_{i=1}^n (v_i/a_i)x_i$, we find the transformed operator equation

$$\tilde{L}w = - \sum_{i=1}^n a_i w_{x_i} + \left(c + \frac{1}{4} \sum_{i=1}^n \frac{v_i^2}{a_i} \right) w = f \exp \left(-\frac{1}{2} \sum_{i=1}^n \frac{v_i}{a_i} x_i \right) \quad (91)$$

Similarly, if c is also constant, for a uniform mesh, the central difference matrix for (90) takes the form

$$A_h = \begin{bmatrix} A & b_1 I & & 0 \\ c_1 I & A & b_2 I & \\ & c_2 & A & \ddots \\ 0 & & \ddots & A \\ & & & c_{n-1} & A \end{bmatrix}$$

If $v_i h/(2a_i) < 1$, then the scheme is of positive type and, as shown in Axelsson (1994), the matrix A_h can be transformed by a diagonal transformation with the block diagonal matrix $D = \text{diag}(d_i D)$, $d_i = 1$, $d_{i+1} = (\sqrt{c_i/b_i})d_i$, $i = 1, \dots, n-1$, so $D^{-1}A_h D$ becomes symmetric when with a similarity transformation $D^{-1}AD$ is made symmetric. Here, the matrix D has a similar form as D but with A, I replaced by scalars. However, when $a_i \ll v_i$, the transformation (91) is not recommended as the problem becomes singularly perturbed and the solution has strong boundary layers.

6.3 Discretization errors

The technique in Section 3.4 can also be applied for certain discretizations of convection-diffusion problems.

Using a proper weight function, it is shown below that the discretization error near Dirichlet boundaries can have a higher order of accuracy than in the interior of the domain. This result is particularly useful in a boundary layer, where the solution is less smooth as its derivatives increase as some power of the Reynolds number ($|v|/\nu$). The result shows that we can balance this growth better, with no need for choosing a much smaller mesh size than is required for stability reasons.

The latter can be illustrated by a simple example.

Example 2. Consider problem (87) with $v = 1$. We assume that $h < 2\varepsilon$ and use two slightly different discretizations:

- central difference approximations everywhere for both terms in (88);
- central difference approximations for the first term everywhere and for the second term in the interval

$(0, 1 - mh)$, but upwind differences for the latter term in the interval $[1 - mh, 1]$. Here, m is a fixed natural number $1 \leq m \leq n$. In particular, this is done for the last interior point.

We show that in the first case the local discretization error has a higher order of approximation near Dirichlet boundaries, and, in the second case, the lower (first) order of approximation near $x = 1$ does not destroy the global (second) order of approximation.

In both cases, the difference operator L_h is of positive type and it is readily seen that there exists a barrier function to L_h , whence L_h is monotone.

The corresponding matrix is tridiagonal,

$$A = \frac{\varepsilon}{h^2} T(-r_i, s_i, -t_i)$$

where $r_i = 1 + h/2\varepsilon$, $s_i = 2$, $t_i = 1 - h/2\varepsilon$, $i = 1, 2, \dots, n$ in case (i) and the same coefficients hold for $i = 1, 2, \dots, n-m$ but $r_i = 1 + h/\varepsilon$, $s_i = 2 + h/\varepsilon$, $t_i = 1$ for $i = n-m+1, \dots, n$ in case (ii). Taking $v = x$ as a barrier function, we obtain $(A_h v)_i = 1$, except for the last point, where

$$(A_h v)_i = \begin{cases} 1 + \frac{1}{2h} \left(\frac{2\varepsilon}{h} - 1 \right) & \text{case (i)} \\ 1 + \frac{\varepsilon}{h^2} & \text{case (ii)} \end{cases}$$

By the Barrier lemma it holds, therefore, that

$$\|A_h^{-1}\|_{\infty} \leq \max_i v_i = 1 \quad (92)$$

For the pointwise discretization error $e_h = u - u_h$, there holds $A_h e_h = \tau_h$, where $\tau_h(x_i) = (L_h u - f_h)(x_i)$ is the (pointwise) truncation error and

$$\|e_h\|_{\infty} \leq \|A_h^{-1}\|_{\infty} \|\tau_h\|_{\infty} \leq \|\tau_h\|_{\infty}$$

More accurate bounds near Dirichlet boundaries are derived next. For this purpose, we partition the matrix A_h in two-by-two block form, $A_h = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$. Here, A_{22} has order $m \times m$ and corresponds to the m last meshpoints. Let the discretization and truncation errors be partitioned consistently with the matrix, $e_h = [e_h^{(1)} \ e_h^{(2)}]^T$ and $\tau_h = [\tau_h^{(1)} \ \tau_h^{(2)}]^T$. From $A_h e_h = \tau_h$, taking the inverse of the partitioned matrix, we obtain

$$\begin{bmatrix} e_h^{(1)} \\ e_h^{(2)} \end{bmatrix} = \begin{bmatrix} (A_{11}^{-1} + A_{11}^{-1} A_{12} S^{-1} A_{21} A_{11}^{-1}) \tau_h^{(1)} \\ + A_{11}^{-1} A_{12} S^{-1} \tau_h^{(2)} \\ S^{-1} A_{21} A_{11}^{-1} \tau_h^{(1)} + S^{-1} A_{22} \tau_h^{(2)} \end{bmatrix} \quad (93)$$

where $S = A_{22} - A_{21} A_{11}^{-1} A_{12}$. The matrix A_{22} has the form

$$A_{22} = \frac{\varepsilon}{h^2} T \left(- \left(1 + \frac{h}{\varepsilon} \right), 2 + \frac{h}{\varepsilon}, -1 \right)$$

while $S = A_{22}$, except its first diagonal entry, which equals $1 + \delta$ for some positive δ , $\delta < 1 + h/\varepsilon$.

Assuming that the solution is sufficiently regular, for the truncation errors, it holds

$$\begin{aligned} \|\tau_h^{(1)}\|_{\infty} &\sim h^2 \left[\frac{1}{12} \varepsilon \|u_x^{(4)}\|_{\infty} + \frac{1}{6} \|u_x^{(3)}\|_{\infty} \right] \\ \|\tau_h^{(2)}\|_{\infty} &= \frac{h^2}{12} \varepsilon \|u_x^{(4)}\|_{\infty} + \frac{h}{2} \|u_x^{(2)}\|_{\infty} \end{aligned} \quad (94)$$

To bound the norms $\|e_h^{(i)}\|_{\infty}$, $i = 1, 2$, we must estimate $\|A_{11}^{-1}\|_{\infty}$ and $\|S^{-1}\|_{\infty}$.

It follows from (92) that $\|A_{11}^{-1}\|_{\infty} \leq 1$. To estimate $\|S^{-1}\|_{\infty}$, we let $\delta = 0$, which will give an upper bound of the norm.

The corresponding matrix \tilde{S} then takes the form

$$\tilde{S} = \frac{\varepsilon}{h^2} \begin{bmatrix} 1 & -1 & & 0 \\ -1 - \sigma & 2 + \sigma & -1 & \\ & \ddots & \ddots & \ddots \\ 0 & & -1 - \sigma & 2 + \sigma \end{bmatrix}$$

where $\sigma = h/\varepsilon$. The matrix \tilde{S} can be factorized as

$$\tilde{S} = \frac{\varepsilon}{h^2} \begin{bmatrix} 1 & & 0 \\ -\theta & 1 & \\ & \ddots & \ddots \\ 0 & & -\theta & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & & 0 \\ & 1 & -1 & \\ & & \ddots & \ddots \\ 0 & & & 1 \end{bmatrix}$$

where $\theta = 1 + \sigma$. Hence,

$$\tilde{S}^{-1} = \frac{h^2}{\varepsilon} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ & \ddots & \ddots & \ddots \\ 0 & & 1 & \theta^{m-1} \end{bmatrix} \begin{bmatrix} 1 & & 0 \\ \theta & 1 & \\ & \ddots & \ddots \\ \theta^{m-1} & & \theta & 1 \end{bmatrix}$$

and

$$\begin{aligned} \tilde{S}^{-1} e &= \frac{h^2}{\varepsilon} \frac{1}{\theta - 1} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ & \ddots & \ddots & \ddots \\ 0 & & 1 & \theta^{m-1} \end{bmatrix} \begin{bmatrix} \theta - 1 \\ \theta^2 - 1 \\ \vdots \\ \theta^m - 1 \end{bmatrix} \\ &= \frac{h^2}{\varepsilon} \frac{1}{(\theta - 1)^2} \begin{bmatrix} \theta^{m+1} - \theta - m(\theta - 1) \\ \theta^{m+1} - \theta^2 - (m-1)(\theta - 1) \\ \vdots \\ (\theta^m - 1)(\theta - 1) \end{bmatrix} \end{aligned}$$

Hence, it can be seen that

$$\begin{aligned} \|S^{-1}\|_{\infty} &\leq \|\tilde{S}^{-1}\|_{\infty} = \frac{h^2}{\varepsilon} \frac{1}{(\theta - 1)^2} [\theta^{m+1} - \theta - m(\theta - 1)] \\ &\leq \frac{h^2}{\varepsilon} \frac{1}{\sigma^2} [e^{(m+1)\sigma} - 1 - (m+1)\sigma] \\ &\approx \frac{1}{2} (m+1) \frac{h^2}{\varepsilon}, \quad \sigma \rightarrow 0 \end{aligned} \quad (95)$$

Since m is fixed, it follows that the quantity $(\varepsilon/h^2) \|S^{-1}\|_{\infty}$ is bounded uniformly in h and ε .

It follows from (93) that $S e_h^{(2)} = \tau_h^{(2)} + A_{21} A_{11}^{-1} \tau_h^{(1)}$. Here, $A_{21} A_{11}^{-1} \tau_h^{(1)} = (1 + \sigma)(A_{11}^{-1} \tau_h^{(1)})_{n-m+1}, \dots, 0]^T$ and by (94) $\|A_{21} A_{11}^{-1} \tau_h^{(1)}\| \leq (1 + \sigma) \|\tau_h^{(1)}\|_{\infty} = O(h^2)$.

Therefore,

$$\|e_h^{(2)}\|_{\infty} \leq \|S^{-1}\|_{\infty} \|\tau_h^{(2)}\|_{\infty} + (1 + \sigma) \|\tau_h^{(1)}\|_{\infty}$$

In case (i) we have $\|\tau_h^{(2)}\|_{\infty} = O(h^2)$, so by (95) $\|e_h^{(2)}\|_{\infty} \leq (1/\varepsilon) O(h^4)$. In case (ii), $\|\tau_h^{(2)}\|_{\infty} = O(h)$ and $\|e_h^{(2)}\|_{\infty} \leq (1/\varepsilon) O(h^3)$. For $e_h^{(1)}$, it follows from (93) that

$$\|e_h^{(1)}\|_{\infty} \leq O(\|\tau_h^{(1)}\|_{\infty} + \|S^{-1}\|_{\infty} \|\tau_h^{(2)}\|_{\infty}) = O(h^2)$$

It is hence seen that for a fixed ε , we have a higher order of convergence at meshpoints in the vicinity of the boundary where Dirichlet boundary conditions are imposed.

In the above estimates, we have not included the dependence of the solution u and its derivatives. Since u has a layer at $x = 1$, it holds that $\|u_x^{(k)}\|_{\infty} = O(\varepsilon^{-k})$. Therefore,

$$\|e_h^{(2)}\|_{\infty} \leq \begin{cases} h^4 \varepsilon^{-4} & \text{case (i)} \\ h^4 \varepsilon^{-4} + h^3 \varepsilon^{-3} = \left(\frac{h}{\varepsilon} \right)^3 \left[1 + \frac{h}{\varepsilon} \right] & \text{case (ii)} \end{cases}$$

To balance this growth, we must use a finer mesh in the layer. Letting h_0 be the mesh size in the domain where the solution is smooth, we then let the mesh size h in the layer domain satisfy

$$\begin{cases} \left(\frac{h}{\varepsilon} \right)^4 = h_0^2, & \text{i.e., } h = h_0^{1/2} \varepsilon, & \text{in case (i)} \\ \left(\frac{h}{\varepsilon} \right)^3 = h_0^2, & \text{i.e., } h = h_0^{2/3} \varepsilon, & \text{in case (ii)} \end{cases} \quad (96)$$

Since the layer term decays exponentially like $\exp(-(1-x)/\varepsilon)$ away from the layer point, a geometrically varying mesh size can be used near this point. The estimates in (96) show that we need a smaller mesh size in case (ii) than in case (i). However, the upwind method has a more

robust behavior for more general problems than the central difference method, so it may still be preferred.

As we have seen, the upwind method is equivalent to an artificial diffusion method with added diffusion of $O(h)$. However, since $h < \varepsilon$, or even $h \ll \varepsilon$ in the layer domain, the resulting smearing of the front is negligible.

Finally, we remark that the central difference method can be used in the domain where the solution is smooth even when $h \geq 2\varepsilon$. The nonphysical oscillations that can arise, do so only if the solution is less smooth, such as in the layer. In practice, it suffices, therefore, to let $h < 2\varepsilon$ there, which is of practical importance when ε is very small.

Example 3. Consider now a two (or higher)-dimensional convection-diffusion problem

$$\mathcal{L}u \equiv -\varepsilon \Delta u + \mathbf{v} \cdot \nabla u + cu = f, \quad \mathbf{x} \in \Omega \subset \mathbb{R}^2 \quad (97)$$

where, for simplicity, $u = 0$ on $\partial\Omega$. In general, $\partial\Omega$ is curvilinear. Let Ω be embedded into a rectangular domain Ω^0 , where we introduce an $n \times m$ uniform rectangular mesh Ω_h^0 , assuming that $n = O(h^{-1})$, $m = O(h^{-1})$. By Ω_h , we denote the mesh subdomain corresponding to the original domain Ω and by Ω_h' , the subset of Ω_h for which there exists a complete difference stencil. For nodes $P \in \Omega_h'$, we use a standard five-point second-order central difference approximation. For nodes in $\Omega_h \setminus \Omega_h'$, we use either a linear interpolation $u_h(P) = [h_E u(E) + h_W u(W)] / (h_E + h_W)$ or a Shortley-Weller difference approximation (21), and a corresponding approximation for the first-order derivative term, where the notations are explained in Figure 2.

Alternatively, as in Example 2, one can use upwind differences for the first-order terms. Similarly, as in Example 2, we can partition the matrix according to the nodeset Ω_h^0 consisting of points near the Dirichlet boundary points, including the outflow boundary (where $v_1 n_1 + v_2 n_2 > 0$ and n_1, n_2 are the components of the outward-pointing normal vector). After somewhat tedious computations in a similar way as in Example 2, one can show that the local discretization errors at meshpoints in Ω_h' have one unit higher order of approximation for the case of a Shortley-Weller approximation and the same order as the global error for the linear approximation. The global error is $O(h^2)$.

The same higher-order approximation holds even if we use an upwind approximation at the outflow part of $\partial\Omega_h$. Finally, similar considerations hold with respect to the ε -dependence as in Example 2.

6.4 Defect correction and adaptive refinement

The previous error estimate results are inapplicable unless L_h is a monotone operator. Furthermore, there can occur

cases where the truncation error does not reveal the real order of the discretization error, as we have seen.

In such cases, a defect-correction method may be useful. This method involves two discrete operators: $L_h^{(1)}$, which is monotone but normally of lower (first) order, as correction operator and a defect operator $L_h^{(0)}$, which may be nonmonotone but is of higher order. The method can be used repeatedly (p times) to achieve a p th order of accuracy.

The corrections to the solution, produced by the defect-correction method, can be used as estimates of the local error to indicate where a mesh refinement should be done. However, we show that the method also allows more accurate error estimates, which are useful in avoiding overrefinement of the mesh. The described mesh-refinement method is special as no slave nodes appear.

Defect-correction methods have been presented in Axelsson and Layton (1990) and Axelsson and Nikolova (1998), among others.

We illustrate the method here on the problem (97), where $\Omega = [0, 1]^2$, $u = g_1$, $\mathbf{x} \in \Gamma_1$, $\mathbf{n} \cdot \nabla u = g_2$, $\mathbf{x} \in \Gamma_2$. Here, $\partial\Omega \equiv \Gamma$ consists of axis-parallel pieces, allowing the use of orthogonal meshes Ω_h . Triangular domains with two axis-parallel edges can be handled in the same way.

We assume that $0 < \varepsilon \leq 1$ (normally $\varepsilon \ll 1$), $c \geq 0$ and the functions f, g_1, g_2 are sufficiently regular. Further, the velocity vector \mathbf{v} is assumed to be sufficiently smooth, $|\mathbf{v}|$ is bounded and the outflow boundary $\Gamma_- = \{\mathbf{x} \in \Gamma; \mathbf{v} \cdot \mathbf{n} < 0\}$ is a subset of Γ_1 .

For this problem, we choose the correction and the defect operators $L_h^{(1)}$ and $L_h^{(0)}$ in the following way:

- $L_h^{(1)}$ is a difference operator of combined central difference and upwind type, in which u_x and u_y are approximated by upwind differences.
- $L_h^{(0)}$ is the second-order central difference operator in which Δu , u_x and u_y are approximated by central differences.

We assume that the local Peclet number $Pe = (2\varepsilon/h|\mathbf{v}|) < 1$; otherwise, the use of upwind discretizations makes no sense.

The second-order defect-correction method includes the following steps:

- (a) Compute an initial approximation $(u_h^{(1)})$ by solving $L_h^{(1)} u_h^{(1)} = f_h$.
- (b) Compute a correction (δ_h) by solving $L_h^{(1)} \delta_h = f_h - L_h^{(0)} u_h^{(1)}$ and set $u_h = u_h^{(1)} + \delta_h$.

Here, f_h denotes the restriction of f and the boundary terms to Ω_h .

In theory (apart from practical implementation details), by performing a sufficient number of defect-correction steps, the defect-correction method can be extended to a p th order difference method. Each application of a correction step increases the order of the error by 1. However, for the considered problem, the error estimates involve powers of ε^{-1} and when $\varepsilon \rightarrow 0$, the above order does not show up unless the local stepsize h is sufficiently small in the subregions where the solution gradient is large. For this purpose, we couple the method with an adaptive refinement method where the mesh size is decreased locally, until some error estimator indicates that the discretization error is sufficiently small.

To simplify the presentation, we assume that the difference mesh consists of square elements, that is, the mesh size used at a meshpoint (x_i, y_j) satisfies $(h_{ij})_x = (h_{ij})_y = h_{ij}$. The mesh size will be refined along an interface and on one side thereof. To be specific, assume that the mesh refinement takes place in a subdomain, as shown in Figure 9. We want to avoid the introduction of the so-called *slave nodes*, because in order to obtain the value of the solution in such a point, an interpolation method of higher order must be used (see Figure 9(a)). Following Axelsson and Nikolova

(1998), we introduce skew-oriented stencils at the interface points, which are not at the center of a cross-oriented stencil (see Figure 9(b)).

The difference approximation in those points will then be based on the equation (97), transformed to a local skew-oriented coordinate system ξ, η , where $\xi = [(\sqrt{2}/2)(x+y)]$ and $\eta = [(\sqrt{2}/2)(x-y)]$. Then, $u_x = [(\sqrt{2}/2)(u_\xi + u_\eta)]$, $u_y = [(\sqrt{2}/2)(u_\xi - u_\eta)]$, and $u_{xx} + u_{yy} = u_{\xi\xi} + u_{\eta\eta}$. Thus, equation (97) takes the form

$$-\varepsilon(u_{\xi\xi} + u_{\eta\eta}) + \frac{\sqrt{2}}{2}(v_1 + v_2)u_\xi + \frac{\sqrt{2}}{2}(v_1 - v_2)u_\eta + cu = f$$

The truncation error of the second-order defect-correction method is

$$\begin{aligned} \tau_h &= L_h^{(1)}(u - u_h^{(1)}) + L_h^{(0)}(u_h^{(1)} - u_h) \\ &= L_h^{(1)}(u - u_h^{(1)}) + L_h^{(0)} u_h^{(1)} - f_h \\ &= (L_h^{(1)} - L_h^{(0)})(u - u_h^{(1)}) + (L_h^{(0)} - \mathcal{L})u \end{aligned}$$

that is, it is a sum of a term arising from the defect-correction method and of the truncation error of the central difference approximation. For the latter term, we have

$$(L_h^{(0)} - \mathcal{L})u_{ij} = \begin{cases} \tau_h^{(0)}(x_i, y_j) & \text{if } (x_i, y_j) \text{ is a regular point} \\ \tau_h^{(0)}(x_i, y_j) - \frac{h_{ij}^2}{2} & \text{if } (x_i, y_j) \text{ is a skew stencil interface point} \\ \times \left(v_2 \frac{\partial^3 u}{\partial x^2 \partial y} + v_1 \frac{\partial^3 u}{\partial x \partial y^2} \right) & \end{cases} \quad (98)$$

where

$$\begin{aligned} \tau_h^{(0)}(x_i, y_j) &= \frac{h_{ij}^2}{6} \left(v_2 \frac{\partial^3 u}{\partial x^2 \partial y} + v_1 \frac{\partial^3 u}{\partial x \partial y^2} \right)_{ij} \\ &\quad - \frac{h_{ij}^2}{12} \left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right)_{ij} + o(h_{ij}^2), \text{ if } u \in C^4(\bar{\Omega}) \end{aligned}$$

Consider now the term $(L_h^{(1)} - L_h^{(0)})(u - u_h^{(1)})$. Let $L_{h,i}^{(i)}$ and $L_{h,j}^{(j)}$, $i = 0, 1$ denote the parts of the difference operator $L_h^{(i)}$ corresponding to the x - and the y -directions correspondingly. The operator $(L_h^{(1)} - L_h^{(0)}) = (L_{h,x}^{(1)} - L_{h,x}^{(0)}) + (L_{h,y}^{(1)} - L_{h,y}^{(0)})$ applied to an arbitrary smooth (at least twice differentiable) function g gives us the following expressions.

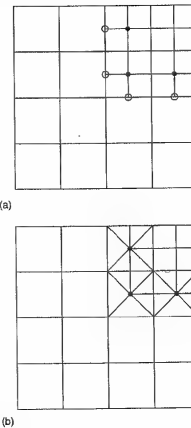


Figure 9. Interface points using (a) slave nodes (b) skewed stencils.

- (1) Let (x_i, y_j) be a regular point. If $v_1 = v_1(x_i, y_j) > 0$, then

$$(L_h^{(1)} - L_h^{(0)})g_{ij} = v_1 \left(\frac{g_{ij} - g_{i-1,j}}{h_{ij}} - \frac{g_{i+1,j} - g_{i,j}}{2h_{ij}} \right) \\ = -\frac{v_1 h_{ij}}{2} \frac{g_{i-1,j} - 2g_{ij} + g_{i+1,j}}{h_{ij}^2} \\ = -\frac{v_1 h_{ij}}{2} \frac{\partial^2 g}{\partial x^2} + o(h_{ij})$$

where the last equality is obtained by Taylor expansion. The $o(h_{ij})$ term stands for a quantity that decays faster than h_{ij} as $h_{ij} \rightarrow 0$. If sufficient regularity holds, its exact expression is

$$\frac{1}{6} \left(\int_{x_{i-1}}^{x_i} (x_{i-1} + s) g^{(4)}(x_i + s) ds \right. \\ \left. + \int_{x_i}^{x_{i+1}} (x_{i+1} - s) g^{(4)}(x_i + s) ds \right)$$

Similar expressions hold for the other Taylor expansions used.

If $v_1 < 0$, we obtain the same relation but with opposite signs. Summing up in both directions, we obtain that for all v_1, v_2

$$(L_h^{(1)} - L_h^{(0)})g_{ij} = -\frac{h_{ij}}{2} \left(|v_1| \frac{\partial}{\partial x^2} + |v_2| \frac{\partial}{\partial y^2} \right) g_{ij} \\ + o(h_{ij}) \quad (99)$$

- (2) Let (x_i, y_j) be an interface point, that is, a skew-oriented stencil is used at this point. If $(v_1 + v_2)_{ij} > 0$, then the truncation error in ξ -direction is

$$(L_h^{(1)} - L_h^{(0)})g_{ij} \\ = \frac{\sqrt{2}}{2} (v_1 + v_2) \left(\frac{g_{ij} - g_{i-1,j-1}}{\sqrt{2}h_{ij}} - \frac{g_{i+1,j+1} - g_{i,j}}{2\sqrt{2}h_{ij}} \right) \\ = -\frac{h_{ij}}{2} (v_1 + v_2) \frac{g_{i-1,j-1} - 2g_{ij} + g_{i+1,j+1}}{(\sqrt{2}h_{ij})^2} \\ = -\frac{h_{ij}}{4} (v_1 + v_2) \left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y} \right)^2 g_{ij} + o(h_{ij})$$

and similar expression holds when $(v_1 + v_2)_{ij} < 0$ and for the velocity direction $v_1 - v_2$. Summing up both directions, for all $v_1 + v_2$ and $v_1 - v_2$, we have

$$(L_h^{(1)} - L_h^{(0)})g_{ij} = -\frac{h_{ij}}{4} \left[|v_1 + v_2| \left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y} \right)^2 \right. \\ \left. + |v_1 - v_2| \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y} \right)^2 \right] g_{ij} + o(h_{ij}) \quad (100)$$

If $v_1 + v_2 > 0$ and $v_1 - v_2 > 0$, the formula simplifies to

$$(L_h^{(1)} - L_h^{(0)})g_{ij} = -\frac{h_{ij}}{2} \\ \times \left[v_1 \left(\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} \right) + 2v_2 \frac{\partial^2 g}{\partial x \partial y} \right] + o(h_{ij})$$

For all v_1, v_2 , the operator $(L_h^{(1)} - L_h^{(0)})g_{ij}$ is a product of h_{ij} , a linear combination of the absolute values of the velocity components and second-order derivatives of g . Thus, the operator is of first-order accuracy if g is smooth.

Lemma 7. Let $u \in C^{(4)}(\bar{\Omega})$ and let $u_h^{(1)}$ be the discrete solution of $L_h^{(1)} u_h^{(1)} = f_h$. Then, there exist functions ϕ and ψ , independent of h_{ij} , such that

$$(u - u_h^{(1)})(x_i, y_j) = h_{ij} \phi(x_i, y_j) + h_{ij}^2 \psi(x_i, y_j) + o(h_{ij}^2) \quad (101)$$

Proof. We introduce the following notations for operators D_i , $i = 2, 3, 4$:

$$D_2 u = v_1 \frac{\partial^2 u}{\partial x^2} + v_2 \frac{\partial^2 u}{\partial y^2}, \quad D_3 u = v_1 \frac{\partial^3 u}{\partial x^3} + v_2 \frac{\partial^3 u}{\partial y^3} \\ D_4 u = \frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}$$

Let (x_i, y_j) be a regular point in Ω_h and $\mathbf{v} = [v_1, v_2] > [0, 0]$. Then, using a Taylor expansion, we get

$$(L_h^{(1)} u - L_h^{(0)} u)_{ij} = -\frac{h_{ij}}{2} D_2 u + \frac{h_{ij}^2}{6} D_3 u \\ - \frac{h_{ij}^3}{12} D_4 u + o(h_{ij}^3) \quad (102)$$

where D_i , $i = 2, 3, 4$ are also taken at the point (x_i, y_j) . Let ϕ and ψ be the solutions of

$$\begin{cases} \mathcal{L}\phi = -\frac{1}{2} D_2 u & \text{in } \Omega \\ \phi = 0 & \text{on } \partial\Omega \end{cases} \quad (103)$$

and

$$\begin{cases} \mathcal{L}\psi = \frac{1}{2} D_2 \phi + \frac{1}{6} D_3 u - \frac{\varepsilon}{12} D_4 u & \text{in } \Omega \\ \psi = 0 & \text{on } \partial\Omega \end{cases}$$

Then, by repeated use of (102), we obtain

$$L_h^{(1)}(u - u_h^{(1)})_{ij} \\ = \left(-\frac{h_{ij}}{2} D_2 u + \frac{h_{ij}^2}{6} D_3 u - \frac{\varepsilon h_{ij}^3}{12} D_4 u \right)_{ij} + o(h_{ij}^2)$$

$$= h_{ij} \mathcal{L}\phi_{ij} + h_{ij}^2 \left(\frac{1}{6} D_3 u - \frac{\varepsilon}{12} D_4 u \right)_{ij} + o(h_{ij}^2) \\ = h_{ij} \mathcal{L}\phi_{ij} + h_{ij}^2 \mathcal{L}\psi_{ij} + o(h_{ij}^2) \\ = h_{ij} L_h^{(1)} \phi_{ij} + h_{ij}^2 L_h^{(1)} \psi_{ij} + o(h_{ij}^2)$$

Since $L_h^{(1)}$ is monotone and its inverse is bounded uniformly in h_{ij} and ε , we find

$$(u - u_h^{(1)})(x_i, y_j) = h_{ij} \phi_{ij} + h_{ij}^2 \psi_{ij} + o(h_{ij}^2)$$

Using Lemma 7 and (96), we obtain now

$$(L_h^{(1)} - L_h^{(0)})(u - u_h^{(1)})_{ij} \\ = -\frac{h_{ij}^2}{2} \left(|v_1| \frac{\partial^2 \phi}{\partial x^2} + |v_2| \frac{\partial^2 \phi}{\partial y^2} \right) + o(h_{ij}^2)$$

for a regular point (x_i, y_j) and a similar expression based on (99) for an interface point. The results for the truncation error are collected in Theorem 18.

Theorem 18. Let u be the solution of (97) and u_h be its discrete solution obtained by applying the second-order defect-correction method. Then, if $u \in C^{(4)}(\bar{\Omega})$ and (x_i, y_j) is a regular point,

$$\tau_h(x_i, y_j) = L_h^{(1)}(u - u_h)_{ij} \\ = (L_h^{(1)} - L_h^{(0)})(u - u_h^{(1)})_{ij} + (L_h^{(0)} - \mathcal{L})u_{ij} \\ = \left[-\frac{1}{2} \left(|v_1| \frac{\partial^2 \phi}{\partial x^2} + |v_2| \frac{\partial^2 \phi}{\partial y^2} \right) + \frac{1}{6} \left(v_1 \frac{\partial^3 u}{\partial x^3} + v_2 \frac{\partial^3 u}{\partial y^3} \right) \right. \\ \left. - \frac{\varepsilon}{12} \left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right) \right] h_{ij}^2 + o(h_{ij}^2) \quad (104)$$

where the function ϕ is the solution of (103). A similar expression holds for interface nodes where (98) and (100) are valid.

Again, using the boundedness of the inverse of $L_h^{(1)}$, we obtain a pointwise estimate of the discretization error, $e_h = u - u_h$.

The result in Theorem 18 remains of limited value unless a proper adjustment of the mesh to the behavior of the solution is made, since the derivatives in (104) are not bounded for $\varepsilon \rightarrow 0$. The latter can be seen from an asymptotic expansion of the solution and its derivatives. Various forms of such expansions have appeared in the literature. A detailed survey in case of parabolic layers is found in Shih and Kellogg (1987) and in the case of exponential layers in Linss and Stynes (1999).

As previously remarked, the errors depend on ε . To obtain a more accurate local estimate of the discretization error than the one that the defect-correction term δ_h provides, one can use the following method, which is based on Lemma 7. From expression (101), it follows that we can compute differences from the current pointwise values of $u_h^{(1)}$ to approximate the derivatives of u with the same (first) order as the approximate values $u_h^{(1)}$. Now, the leading part of the truncation error $L_h^{(1)}(u - u_h^{(1)})$ takes the form

$$\tau_h^{(1)}(x_i, y_j) \\ = \begin{cases} -\frac{h_{ij}}{2} \left(|v_1| \frac{\partial^2 u}{\partial x^2} + |v_2| \frac{\partial^2 u}{\partial y^2} \right)_{ij} & \text{if } (x_i, y_j) \text{ is a regular point} \\ -\frac{h_{ij}}{2} \left(|v_1| \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + 2|v_2| \frac{\partial^2 u}{\partial x \partial y} \right)_{ij} & \text{if } (x_i, y_j) \text{ is a skew-stencil interface point} \end{cases}$$

By solving an additional system $L_h^{(1)} \tau_h^{(1)} = \tau_h^{(1)}$, we can hence estimate the discretization error $u - u_h^{(1)}$. Since $e_h = u - u_h = (u - u_h^{(1)}) - \delta_h$, where δ_h is computed in the defect-correction method, we have an accurate pointwise estimate of $u - u_h$ to be used as an error indicator in an adaptive local mesh-refinement procedure.

The adaptive strategy can hence be to refine a current mesh Ω_h by patching at every point or at the surrounding of points, where the approximation of e_h , $\tau_h^{(1)}$, is larger than some tolerance tol . The tolerance can be an absolute small quantity or a relative quantity, $\text{tol} = \mu \max_{ij} |\tau_h^{(1)}(x_i, y_j)|$, $0 < \mu < 1$. A slight disadvantage of the adaptive refinement method is that it leads to nonsymmetric algebraic systems.

Numerical tests for singularly perturbed problems with the above criterion can be found in Axelsson and Nikolova (1998).

6.5 The time-dependent convection-diffusion problem

The time-dependent convection-diffusion equation can be classified as a parabolic equation with lower-order terms. In one dimension, it has the form

$$u_t = \varepsilon u_{xx} - v u_x - c u + f, \quad x \in [a, b], \quad t > 0 \quad (105)$$

with some proper boundary and initial conditions.

We show below that the lower-order terms may affect stability. Consider (105) with $v > 0$ and $c = 0$, $f = 0$,

$u(a, t) = \alpha$, $u(b, t) = \beta$, and $u(x, 0) = u_0$, using the explicit scheme

$$\frac{u_n^{(m+1)} - u_n^{(m)}}{k} = \varepsilon \frac{u_{n+1}^{(m)} - 2u_n^{(m)} + u_{n-1}^{(m)}}{h^2} + v \frac{u_{n+1}^{(m)} - u_{n-1}^{(m)}}{2h} = 0 \quad (106)$$

The scheme is only conditionally stable, namely, the relation $\varepsilon(k/h^2) \leq (1/2)$ must hold.

Remark 10. Although the scheme (106) has accuracy $O(k + h^2)$, since $k \leq (h^2/2\varepsilon)$, the scheme is second-order accurate.

We set $\mu = (k/h^2)$ (referred to as the Courant number) and $\alpha = (vh/2\varepsilon)$ (the cell Peclet number) and rewrite (106) as

$$u_n^{(m+1)} = (1 - 2\varepsilon\mu)u_n^{(m)} + \varepsilon\mu(1 + \alpha)u_{n-1}^{(m)} + \varepsilon\mu(1 - \alpha)u_{n+1}^{(m)} \quad (107)$$

For parabolic problems, the following maximum principle holds.

$$\sup_x |u(x, t)| \leq \sup_x |u(x, \bar{t})| \quad \text{for } t \leq \bar{t}$$

From (107), one can see that the numerical scheme will satisfy the maximum principle if and only if $\alpha \leq 1$, that is, $h \leq (2\varepsilon/v)$. The latter is automatically satisfied for upwind schemes and acts as a restriction on the spatial mesh for central differences.

6.6 A generalized difference scheme by use of local Green's functions

For some differential operators, the analytic form of their fundamental solution is known. It can be used to construct an exact or approximate local Green's function with homogeneous boundary values at discretization element edges (points).

Let $\mathcal{L}u = -\nabla \cdot (\varepsilon \nabla u) + \mathbf{b} \cdot \nabla u + cu = f$, $\mathbf{x} \in \Omega$, $u = 0$ on $\partial\Omega$ (108)

The operator \mathcal{L}^* adjoint to the operator \mathcal{L} has the following form

$$\mathcal{L}^*v = -\nabla \cdot (\varepsilon \nabla v) - \nabla \cdot (\mathbf{b} v) + cv, \quad \mathbf{x} \in \Omega, v = 0 \text{ on } \partial\Omega$$

(Here ε, \mathbf{b} may be functions of \mathbf{x} .) Let Ω_h be a difference mesh on Ω , not necessarily uniform, and let g_i be the local

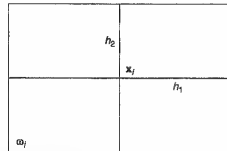


Figure 10. Subdomain ω_i .

Green's function to \mathcal{L}^* at \mathbf{x}_i , that is,

$$\mathcal{L}^*g_i = \delta(\mathbf{x} - \mathbf{x}_i) \quad \text{in } \omega_i, g_i = 0 \text{ on } \partial\omega_i$$

where ω_i is typically an element as shown in Figure 10.

Here, $\delta(\mathbf{x} - \mathbf{x}_i)$ is the Kronecker distribution function at \mathbf{x}_i . If g_i is not known analytically, we assume that a function \tilde{g}_i is known, which is a sufficiently accurate approximation of g_i .

Owing to the homogeneous boundary values of g_i on $\partial\omega_i$, it holds that

$$\begin{aligned} \int_{\omega_i} f \tilde{g}_i d\omega &= \int_{\omega_i} \mathcal{L}u \tilde{g}_i d\omega \\ &= \int_{\omega_i} (\varepsilon \nabla u \cdot \nabla \tilde{g}_i - \nabla \cdot (\mathbf{b} \tilde{g}_i)u + cu \tilde{g}_i) d\omega \\ &= \int_{\omega_i} \mathcal{L}^* \tilde{g}_i u d\omega + \oint_{\partial\omega_i} \varepsilon \frac{\partial \tilde{g}_i}{\partial \mathbf{n}} u d\gamma \\ &= \int_{\omega_i} \mathcal{L}^* g_i u d\omega + \oint_{\partial\omega_i} \varepsilon \frac{\partial \tilde{g}_i}{\partial \mathbf{n}} u d\gamma + \int_{\omega_i} \mathcal{L}^* (\tilde{g}_i - g_i) u d\omega \\ &= u(\mathbf{x}_i) + \oint_{\partial\omega_i} \varepsilon \frac{\partial \tilde{g}_i}{\partial \mathbf{n}} u d\gamma + \int_{\omega_i} \mathcal{L}^* (\tilde{g}_i - g_i) u d\omega \quad (109) \end{aligned}$$

For a one-dimensional problem, the local Green's function at \mathbf{x}_i satisfies

$$\begin{aligned} \mathcal{L}^*g_i(x) &= 0, \quad x \in (x_{i-1}, x_i) \cup (x_i, x_{i+1}) \\ g_i(x) &= 0, \quad x \in [0, x_{i-1}] \cup [x_{i+1}, 1] \\ \varepsilon(x_i)[g_i'(x_i-) - g_i'(x_i+)] &= 1 \quad (110) \end{aligned}$$

The generalized difference scheme (109) with $\tilde{g}_i = g_i$ then takes the form of a three-point difference method

$$\begin{aligned} -\varepsilon_{i-1}g_i'(x_{i-1})u_{i-1} + u_i + \varepsilon_{i+1}g_i'(x_{i+1})u_{i+1} \\ = \int_{x_{i-1}}^{x_{i+1}} f g_i dx, \quad i = 1, 2, \dots, n \quad (111) \end{aligned}$$

where $\{x_i\}_{i=0}^{n+1}$ is a division of the interval $[0, 1]$.

For the operator $-u_{xx} + bu_x = f$, $0 < x < 1$, $u(0) = u(1) = 0$, letting $b = b(x)$ be constant when evaluating g_i and assuming for simplicity a uniform mesh, formula (111) takes the form

$$-\frac{1}{1 + e^{-bh}}u_{i-1} + u_i - \frac{1}{1 + e^{bh}}u_{i+1} = \int_{x_{i-1}}^{x_{i+1}} f g_i dx, \quad i = 1, 2, \dots, n \quad (112)$$

A computation shows that

$$g_i = \frac{1}{b_i(1 + e^{bh})} \begin{cases} e^{bh} - e^{b(x_i-x)}, & x_{i-1} < x < x_i \\ e^{b(x_{i+1}-x)} - 1, & x_i < x < x_{i+1} \end{cases} \quad (113)$$

When $b_i \rightarrow 0$,

$$g_i \rightarrow \tilde{g}_i = \frac{1}{2} \begin{cases} x - x_{i-1}, & x_{i-1} < x < x_i \\ x_{i+1} - x, & x_i < x < x_{i+1} \end{cases}$$

which reduces (111) to the central difference method except for the right side function. When b is constant and $\int_{x_{i-1}}^{x_{i+1}} f \tilde{g}_i dx$ is evaluated exactly, the corresponding difference method gives the exact solution at the nodepoints.

Remark 11. The functions \tilde{g}_i are the standard 'hat' functions used as basis functions in linear finite element methods.

When solving (109), we can use a defect-correction method, that is, first compute an initial approximation $u^{(0)}$ from

$$u^{(0)}(x_i) + \oint_{\partial\omega_i} \varepsilon \frac{\partial \tilde{g}_i}{\partial \mathbf{n}} u^{(0)} d\gamma = \int_{\omega_i} f \tilde{g}_i d\omega, \quad i = 1, 2, \dots, n \quad (114)$$

and a correction (once or more times) using

$$\delta u^{(0)}(x_i) + \oint_{\partial\omega_i} \varepsilon \frac{\partial \tilde{g}_i}{\partial \mathbf{n}} \delta u^{(0)} d\gamma = \int_{\omega_i} \mathcal{L}(u - u^{(0)}) \tilde{g}_i d\omega, \quad i = 1, 2, \dots, n \quad (115)$$

Then let $u^{(1)} = u^{(0)} + \delta u^{(0)}$ and repeat if necessary.

Since $(\partial \tilde{g}_i / \partial \mathbf{n}) < 0$ at $\partial\omega_i$, it follows that the matrix A_h in the arising linear systems in (114) and (115) of lowest order is an M -matrix and has a bounded inverse.

In the above method, we have deleted the last term in (109), which contains the unknown exact Green's function g_i . If \tilde{g}_i is a sufficiently accurate approximation of g_i , the convergence of the defect-correction method is fast.

One can determine an accurate approximation by taking the fundamental solution to $\varepsilon \Delta u$ and multiplying it by the 1D local Green's functions to assuming now for simplicity that $c = 0$ $\varepsilon u_{xx} - (b_1 u)_x = \delta(\mathbf{x} - \mathbf{x}_i)$ and $\varepsilon u_{yy} - (b_2 u)_y =$

$\delta(\mathbf{x} - \mathbf{x}_i)$, the solution of which is given in (113), both with homogeneous boundary conditions.

In practice, u is approximated by piecewise polynomials such as in finite element methods. Let its approximation be u_h and let the discretization error be split as $u - u_h = \eta - \theta_h$, where $\eta = u - u_h$, $\theta_h = u_h - u_h$, and u_h is the interpolant in the space spanned by the piecewise polynomials. Applying this error estimate in (114) and (115), and using the triangle inequality $|u - u_h| \leq |\eta| + |\theta_h|$ we readily find the pointwise error bound

$$\max_i |(u - u_h)(x_i)| \leq C \|A_h^{-1}\| \max_i |(u - u_h)(x_i)|$$

In a 1D problem, the interpolation errors at x_i are zero, so the corresponding approximations in (113) using the exact local Green's function takes the exact values $u(x_i)$ at the nodepoints.

6.7 A meshless difference method

On the basis of (109), a generalized difference scheme can be derived, which is even meshless, that is, it is applicable for an arbitrary distribution of nodepoints in Ω and is not based on a difference mesh.

Many important problems in practice, like crack propagation, fragmentation, and large deformations, are characterized by a continuous change in the geometry of the domain under analysis. Conventional finite difference or finite element methods for such problems can be cumbersome and expensive, as they may require continuous remeshing of the domain to avoid breakdown of the calculation due to excessive mesh distortions. Meshless methods provide an attractive alternative for the solution of such classes of problems. For a meshless method not based on (109) and for using nonpolynomial interpolation functions, see Duarte and Oden (1996) and the references stated therein.

Let $\mathcal{L}u = f$ in Ω and $u = g$ on $\partial\Omega$ be given, where Ω is a bounded domain and \mathcal{L} is a second-order elliptic operator. Although the method can be applied for more general cases, where only some (sufficiently accurate) approximate Green's function is known, we consider here the case where the Green's function for the adjoint operator on a disc is known. Further, even though the construction is applicable for more general operators, for simplicity, we consider here only the operator $\mathcal{L} = -(\partial^2/\partial x^2) - (\partial^2/\partial y^2)$.

Let $\{x_1, \dots, x_n\}$ be a set of disjoint points in $\bar{\Omega}$ and let $V = \{x_1, \dots, x_n\}$ be the subset of interior points. For each point x_i in V , we choose a set of $q_i (< n)$ neighboring points $P(x_i) = \{x_i^{(1)}, \dots, x_i^{(q_i)}\}$ such that all angles

$\mathcal{L}(x_i, x_k^{(i)})$ are different. The aim is to derive a local difference approximation at x_i in the form

$$\tilde{u}(x_i) - \sum_{k=1}^m \gamma_k^{(i)} \tilde{u}(x_k^{(i)}) = \int_{\omega_i} f g_i, \quad i = 1, \dots, m \quad (116)$$

where $\tilde{u}(x_i)$ denotes the corresponding approximation to $u(x_i)$ and g_i is the local Green's function, which satisfies $\mathcal{L}^* g_i = \delta(x_i, \cdot)$ in ω_i , $\delta(x_i, \cdot)$ is the Dirac measure at x_i and the trace $tr(g_i)$ of g_i is equal to zero on $\partial\omega_i$. Here, ω_i is a disc with center at x_i , the radius of which will be determined later. It then holds that $g_i = (1/2\pi) \ln(r_i/r)$, $0 < r < r_i$. It is assumed that $\int_{\omega_i} f g_i$ is bounded, otherwise the corresponding singularity of u must be subtracted from the solution.

It is further assumed that the union $\bigcup_{i=1}^m \omega_i$ of the discs cover Ω .

The basic relation to be used is

$$\begin{aligned} \int_{\omega_i} f g_i &= \int_{\omega_i} \mathcal{L} u g_i = \int_{\omega_i} \nabla u \cdot \nabla g_i - \oint_{\partial\omega_i} \frac{\partial u}{\partial \mathbf{n}} g_i \\ &= \int_{\omega_i} \mathcal{L}^* g_i u + \oint_{\partial\omega_i} \frac{\partial g_i}{\partial \mathbf{n}} u \end{aligned}$$

or

$$\int_{\omega_i} f g_i = u(x_i) + \oint_{\partial\omega_i} \frac{\partial g_i}{\partial \mathbf{n}} u \quad (117)$$

Here, u will be approximated by polynomials of degree p_i , $p_i \geq 2$. For computational reasons, hereby it is efficient to use the first harmonic polynomials of \mathcal{L} , but written as trigonometric functions in polar coordinates, complemented with the missing polynomials, which are not harmonic. In polar coordinates,

$$\mathcal{L} = -\frac{\partial^2 u}{\partial r^2} - \frac{1}{r} \frac{\partial u}{\partial r} - \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2}$$

and $r^k \cos k\theta$, $r^k \sin k\theta$, $k = 1, \dots, p_i$ are the first harmonic polynomials.

We note that to model corner singularities, one can use the fractional power form of the latter functions, that is, $r^\alpha \sin \alpha\theta$, where $\alpha = \pi/\omega$ for a corner with interior angle ω , if $\omega \neq \pi/n$, and $r^\alpha \ln r \sin \alpha\theta$ if $\omega = \pi/n$, $n = 1, 2, \dots$

We now approximate u in (117) locally around x_i by a linear combination of those functions (i.e. we use a trigonometric expansion plus possibly some terms corresponding to the missing polynomials up to degree p_i).

Substituting the polynomials in (117) results in a linear system for the corresponding coefficients. It is not necessary to solve this system but the arising matrix can be used for the computation of the coefficients $\gamma_k^{(i)}$, $k = 1, \dots, g_i$ in (116).

For certain regular distributions of nodepoints, it turns out that there is a cancellation of some error terms implying that g_i can be much smaller than the total number of polynomials (monomials) used, for example, in the Pascal triangle, there appear $p_i(p_i + 1)/2$ such monomials.

The radius r_i in ω_i is determined to make the difference approximation in (116) exact for the missing second-order polynomial $u = x^2 + y^2$. Then, $\mathcal{L}u = -4$ and (116) take the form

$$\sum_{k=1}^{p_i} \gamma_k^{(i)} r_k^{(i)} = 4 \int_{\omega_i} g_i d\omega$$

where $r_k^{(i)} = |x_k^{(i)} - x_i|$. From $g_i = (1/2\pi) \ln(r_i/r)$, it follows that

$$\oint_{\partial\omega_i} g_i d\omega = \frac{1}{4} r_i^2$$

so

$$r_i = \left\{ \sum_{k=1}^{p_i} \gamma_k^{(i)} |x_k^{(i)} - x_i|^2 \right\}^{1/2}$$

that is, since $\sum_{k=1}^{p_i} \gamma_k^{(i)} = 1$ (take $u \equiv 1$ in (116)), r_i form a weighted average of distances between the local neighbors.

Discretization error

The exact solution satisfies (117), while the difference approximation is

$$\tilde{u}(x_i) - \sum_{k=1}^{p_i} \gamma_k^{(i)} \tilde{u}(x_k^{(i)}) = R_i(f g_i) \quad (118)$$

Here, $R_i(f g_i)$ denotes the quadrature approximation used for $\int_{\omega_i} f g_i$, with error $\int_{\omega_i} f g_i - R_i(f g_i)$.

Writing equation (117) in the form

$$\begin{aligned} u(x_i) - \sum_{k=1}^{p_i} \gamma_k^{(i)} u(x_k^{(i)}) &= \oint_{\partial\omega_i} \frac{\partial g_i}{\partial \mathbf{n}} u \\ &\quad - \sum_{k=1}^{p_i} \gamma_k^{(i)} u(x_k^{(i)}) + \int_{\omega_i} f g_i \end{aligned}$$

and subtracting equation (118), we obtain

$$e_h(x_i) - \sum_{k=1}^{p_i} \gamma_k^{(i)} e_h(x_k^{(i)}) = \delta_1(u) + \delta_2(u)$$

where $e_h = u - \tilde{u}$ and

$$\delta_1(u) = \oint_{\partial\omega_i} \frac{\partial g_i}{\partial \mathbf{n}} u - \sum_{k=1}^{p_i} \gamma_k^{(i)} u(x_k^{(i)})$$

$$\delta_2(u) = \int_{\omega_i} f g_i - R_i(f g_i)$$

that is, denote the line integral and the domain integral errors, respectively. It is readily seen that the line integral error can be written as a line integral of the interpolation error in u .

Lemma 8. Let u_k denote the interpolation of u on $x_k^{(i)}$, $k = 1, \dots, p_i$. Then,

$$\delta_1(u) = \oint_{\partial\omega_i} \left(-\frac{\partial g_i}{\partial \mathbf{n}} \right) (u - u_h)$$

If $\gamma_k^{(i)} \geq 0$, then the matrix A_h in (116) is an M -matrix. It can be seen that the smallest eigenvalue of the matrix is bounded below by $\min_k O(r_k^2) = h_0^2$ and likewise $\|A_h^{-1}\|_\infty = O(h_0^{-2})$. For the lowest (second)-order schemes, it holds that $\gamma_k^{(i)} \geq 0$. However, the latter does not hold for higher-order schemes, and one can expect that some of the coefficients are negative in general. To solve the resulting linear system with a matrix A_h , it can be efficient to use a defect-correction method with a second-order (or possibly only first order) scheme, for which the corresponding values of $\gamma_k^{(i)} \geq 0$, as corrector for the higher-order scheme.

Theorem 19. Assume that the difference scheme corresponds to exact interpolation of all polynomials of degree p_i at least, and that the solution is sufficiently smooth. Assume also that $\epsilon h \leq r_i \leq \bar{\epsilon} h$ for some constants $\epsilon, \bar{\epsilon}$. Then, using some steps of a defect-correction method, if necessary, the pointwise discretization error satisfies

$$|e_h(x_i)| \leq O(h^{p-1})$$

where $p = \min_i p_i$.

Proof. By Lemma 8, the local discretization (interpolation and integration) errors are $O(r_i^{p+1}) \leq O(h^{p+1})$. The local errors are globally coupled by the assembled matrix A_h , whose inverse is bounded by $O(h^{-2})$. Hence, $\|e_h\|_\infty \leq \|A_h^{-1}\|_\infty O(h^{p+1}) = O(h^{p-1})$. \square

Example 4 [A regular distribution of points] We consider now the case where the nodepoints are chosen as the vertices in a hexagonal mesh, that is, we use a seven-point interpolation scheme. Owing to symmetries of the scheme, it follows readily that $\delta_1(u) = 0$ for all polynomials of degree five in variables x, y . Hence, the scheme is fourth-order accurate if $\int_{\omega_i} f g_i$ is computed with corresponding accuracy.

Similar results hold for the vertices in a cuboctahedral mesh.

6.8 A hybrid method of characteristics and central difference method

As remarked previously, in many diffusion processes arising in physical problems, convection dominates diffusion, and it is natural to seek numerical methods that reflect their almost hyperbolic nature.

The convective character is easily seen by considering the method of characteristics for the reduced equation,

$$\mathbf{v} \cdot \nabla u + cu = f \text{ in } \Omega, \quad u = g \text{ on } \Gamma_- = \{x \in \Gamma, \mathbf{v} \cdot \mathbf{n} < 0\}$$

Let $x(t, s)$ be the parametric representation of the lines of characteristics defined by the vector field through the point (x_0, y_0) on Γ_- , that is, we have $x = x_1(t, s)$, $y = x_2(t, s)$ for points on this line and

$$\begin{aligned} \frac{dx(t, s)}{dt} &= \mathbf{v}(x, y) = \mathbf{v}(x(t, s)), \quad t > 0, \\ \mathbf{z}(0, s) &= (x_0, y_0) \in \Gamma_- \end{aligned}$$

Since the vector field is uniquely defined, no two characteristic lines may cross each other. Using the chain rule, we obtain

$$\frac{d\hat{u}(t)}{dt} = \sum_{i=1}^2 \frac{\partial u}{\partial x_i} \frac{dx_i}{dt} = \mathbf{v} \cdot \nabla u$$

where $\hat{u}(t) = u(x(t, s))$ and s is fixed. Hence,

$$\frac{d\hat{u}(t)}{dt} + c\hat{u} = f(x(t, s)), \quad t > 0, \quad \hat{u}(0) = u(x_0, y_0)$$

so when the characteristic lines have been computed, the solution of the reduced equation along each characteristic line can be computed as the solution of an initial value problem for an ordinary differential equation.

When ϵ is small and $|\mathbf{v}| > 0$, the solution of (108) is close to the solution of the reduced equation except in subdomains where boundary layers occur such as when the solution of the reduced equation does not satisfy the boundary conditions on the outflow boundary.

We describe here a combination of a method of characteristics and a central difference method. The method is illustrated on a 1D problem. For a treatment of 2D or higher-dimensional problems, see Axelsson and Marinova (2002).

The convection-diffusion problem then has the following form:

$$\begin{aligned} \mathcal{L}u &\equiv -\epsilon u'' + vu' + cu = f, \quad 0 < x < 1, \\ u(0) &= 0, \quad u(1) = 1 \end{aligned} \quad (119)$$

We assume that $v \geq v_0 > 0$ and $c \geq 0$ in $[0, 1]$ and that v and c are bounded C^1 functions, $v, c \in C^1[0, 1]$.

The difference scheme

Let $\theta, 0 < \theta < 1$, be a variable weight coefficient. The difference method on an arbitrary mesh $\Omega_N = \{x_i, i = 0, 1, \dots, N, x_0 = 0, x_N = 1, x_i < x_{i+1}\}$ with variable step $h_i = x_i - x_{i-1}, i = 1, \dots, N$ (N is the number of the mesh intervals) takes the form

$$\begin{aligned} L^N u_i^N = & -\frac{2\varepsilon}{h_i + h_{i+1}} \left(\frac{u_{i+1}^N - u_i^N}{h_{i+1}} - \frac{u_i^N - u_{i-1}^N}{h_i} \right) + \theta_i \\ & \times \left[v(x_i) \frac{u_{i+1}^N - u_{i-1}^N}{h_i + h_{i+1}} + c(x_i) u_i^N \right] + (1 - \theta_i) \\ & \times \left[v \left(x_i - \frac{h_i}{2} \right) \frac{u_i^N - u_{i-1}^N}{h_i} + c \left(x_i - \frac{h_i}{2} \right) \frac{u_i^N + u_{i-1}^N}{2} \right] \\ = & f^N = \theta_i f(x_i) + (1 - \theta_i) f \left(x_i - \frac{h_i}{2} \right) \end{aligned}$$

for each interior mesh point $x_i, i = 1, 2, \dots, N-1$. Here,

- u_i^N denotes the finite difference approximation of the solution u at mesh point $x_i, i = 1, 2, \dots, N-1$;
- $\theta_i = [1/(1+r_i)]$, where $r_i = [v(x_i)h/2\varepsilon]$ is the local Peclet number, $h = \max\{h_i, h_{i+1}\}$;
- $1 - \theta_i = [r_i/(1+r_i)] = [v(x_i)h/2\varepsilon + v(x_i)h]$.

The corresponding finite difference mesh is illustrated in Figure 11.

The scheme is a linear combination of a central difference scheme at x_i and at $x_i - (h_i/2)$, except that the central difference for the second-order derivative is evaluated only at x_i . The scheme is a three-point upwind scheme. When $\varepsilon \ll h$, it is dominated by the approximation at $x_i - (h_i/2)$, while when $\varepsilon \gg h$, it is dominated by the central difference approximation at x_i .

We assume that the mesh is uniform or varies smoothly when $\varepsilon \geq h$.

It is readily seen that the operator L^N is monotone if $h_i \leq 2v_0/\max_{x \in \Omega_N} c(x)$.

With a barrier function $w = x|_{\Omega_N}$, a straightforward computation shows that

$$L^N w \geq v_0 e, \quad e = (1, 1, \dots, 1)^T \quad (120)$$

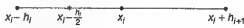


Figure 11. Finite difference mesh in a 1D case.

so by the Barrier Lemma,

$$\|(L^N)^{-1}\| \leq \frac{1}{v_0} \|w\| = \frac{1}{v_0} \quad (121)$$

which holds uniformly in ε .

Truncation error estimate

To estimate the discretization error $\|u - u^N\|$, we will first prove the boundedness of the truncation error $L^N[u(x_i) - u_i^N] = L^N u(x_i) - f^N$. Assuming first $h = h_i = h_{i+1}$ and using a Taylor expansion and by a straightforward derivation, rearranging terms (see Axelsson and Nikolova, 1998), we obtain the following result for the truncation error,

$$\begin{aligned} L^N u(x_i) - f^N = & -\varepsilon(1 - \theta_i) \left[u''(x_i) - u'' \left(x_i - \frac{h}{2} \right) \right] \\ & - \frac{1}{12} \varepsilon h^2 u^{(iv)} + \frac{1}{6} \theta_i v(x_i) h^2 u'' \\ & + (1 - \theta_i) \left[\frac{1}{24} v \left(x_i - \frac{h}{2} \right) h^2 u'' + \frac{1}{8} c \left(x_i - \frac{h}{2} \right) h^2 u'' \right] \end{aligned}$$

By the choice of θ_i , the error arising from the term

$$-\varepsilon(1 - \theta_i) \left[u''(x_i) - u'' \left(x_i - \frac{h}{2} \right) \right]$$

is

$$\begin{aligned} -\frac{\varepsilon r_i}{1+r_i} \left[u''(x_i) - u'' \left(x_i - \frac{h}{2} \right) \right] = & -\frac{\varepsilon v(x_i) h}{2\varepsilon + v(x_i) h} \frac{h}{2} u'' \\ \leq & \frac{1}{2} \min\{v(x_i) h, 2\varepsilon\} h |u''| \leq \frac{1}{2} \max_{x \in [0,1]} v(x) h^2 |u''| \end{aligned}$$

In the above equation, the derivatives $u'', u''', u^{(iv)}$ are evaluated at some intermediate points. With an integral form of the remainder, we obtain the following estimates for the truncation error:

$$\begin{aligned} |L^N u(x_i) - f_i^N| \leq & C_1 \varepsilon h \int_{x_{i-1}}^{x_{i+1}} |u^{(iv)}(\xi)| d\xi + C_2 \\ & \times h \int_{x_{i-1}}^{x_{i+1}} |u''(\xi)| d\xi + C_3 h \int_{x_{i-1}}^{x_{i+1}} |u''(\xi)| d\xi \quad (122) \end{aligned}$$

where C_1, C_2, C_3 are positive constants.

If the mesh is irregular and $h_i \neq h_{i+1}$, then for the estimate of the truncation error at x_i , we have

$$\begin{aligned} |L^N u(x_i) - f_i^N| \leq & C_1 \varepsilon \int_{x_{i-1}}^{x_{i+1}} |u^{(iv)}(\xi)| d\xi \\ & + C_2 \int_{x_{i-1}}^{x_{i+1}} |u''(\xi)| d\xi + C_3 \int_{x_{i-1}}^{x_{i+1}} |u''(\xi)| d\xi \quad (123) \end{aligned}$$

where $\bar{h}_i = \max\{h_i, h_{i+1}\}$.

Discretization error estimate

The estimate of the discretization error includes the following steps. First, an estimate of the truncation error, second, a construction of a suitable barrier function w^N , and, finally, an estimate of the discretization error based on a discrete comparison principle.

To resolve the solution in the layer, one can use a mesh refinement – a graded or a uniform mesh in the layer part. Consider the convection-dominated case $\varepsilon \ll N^{-1}$. Set $\tau = \min\{(1/2), (2/\beta)\varepsilon \ln N\}$, where $0 < \beta < v_0$, and denote

$$\begin{aligned} \Omega_1^N = & \left\{ i \frac{1-\tau}{N}, i = 0, \dots, \frac{N}{2} \right\} \\ \Omega_2^N = & \left\{ x_i, i = \frac{N}{2+1}, \dots, N \right\} \end{aligned}$$

Then $\Omega^N = \Omega_1^N \cup \Omega_2^N$. The mesh points x_i in Ω_2^N are defined by

- $x_i = 1 - \tau + \tau(i - N/2)/(N/2)$ for the piecewise uniform mesh;
- $x_i = 1 - \tau + \tau \ln(i + 1 - N/2)/\ln(N/2 + 1)$ for the logarithmically graded mesh.

We use the notation Ω_θ for the uniform mesh, called Shishkin mesh (Shishkin, 1990) and Ω_τ for the logarithmically graded mesh, which is of Bakhvalov type (Bakhvalov, 1969).

We consider here a corresponding mesh-refinement method. To analyze the method, we use the following splitting of the solution:

$$u = g + z$$

The smooth part g satisfies the problem

$$Lg = f, \quad g(0) = u(0) = 0, \quad v(1)g'(1) + c(1)g(1) = f(1)$$

It is assumed that the given data is such that

$$|g^{(k)}| \leq C \quad \text{for } 0 \leq k \leq 4$$

The layer part z satisfies

$$Lz = 0, \quad z(0) = 0, \quad z(1) = u(1) - g(1)$$

and

$$|z^{(k)}| \leq C \varepsilon^{-k} \exp \left(-\frac{\beta(1-x)}{\varepsilon} \right) \quad \text{for } 0 \leq k \leq 4 \quad (124)$$

By considering each term of the right-hand side of the following inequality separately,

$$\|u - u^N\| \leq \|g - g^N\| + \|z - z^N\|$$

where g^N, z^N are the corresponding finite difference approximations to g and z respectively, the following discretization error estimate can be proven.

Theorem 20. Let u be the solution of the convection-diffusion problem (119) and u^N be its discrete solution obtained by applying the second-order hybrid method on the Shishkin mesh Ω_θ . Let $v, c, f \in C^2[0, 1]$ and $v \geq v_0 > \beta > 0$. Then, the discretization error is globally uniformly bounded in ε by

$$\|u - u^N\| \leq CN^{-2}(\ln N)^2 \quad (125)$$

where N is the total number of points used.

Remark 12. In a similar way, one can show for the exponentially graded mesh a uniform optimal order estimate $\|u - u^N\| \leq CN^{-2}$ (see Axelsson and Nikolova, 1998 for details of the technique to be used). As shown in Axelsson and Marinova (2002), the above results are nicely illustrated from numerical tests.

Remark 13. Instead of the present choice of $\theta, \theta = 2\varepsilon/(2\varepsilon + v|h|)$, one can let $\theta = \min\{1, 2\varepsilon/(|v|h)\}$. Depending on the relative signs of the discretization error terms, one of the two choices may give slightly smaller errors than the other.

Remark 14. For a combined method of characteristics and a Galerkin finite element method applied for time-dependent problems, see Douglas and Russell (1982).

Final remarks: Difference methods are most useful for regular or quasi-regular meshes where the mesh size varies smoothly, such as $h_{i+1/2} = [1 + O(h_{ij})]h_{ij}$. Otherwise, a defect-correction method can be applied.

Alternative methods for irregular meshes can be based on finite element or finite volume methods. It can often be efficient to write a second-order problem as a system of equations of first order, for instance, when the coefficient function a varies strongly locally. For instance, the diffusion problem $-\nabla(a\nabla u) + cu = f$ can be written in the form

$$\begin{cases} \frac{1}{a} z - \nabla u = 0 \\ \nabla \cdot z - cu = -f \end{cases} \quad (126)$$

and discretized by finite element methods that satisfy a certain *inf-sup* stability condition; see, Brezzi and Fortin (1991).

As shown in Axelsson and Gustafsson (1991), if one uses a piecewise constant approximation for z_1, z_2 and piecewise linear for u in (126), then using the simplest numerical quadrature for $\int(1/a)$, the method reduces to the standard finite element method, or equivalently, the

five-point difference method. For variable coefficient a and higher-order quadrature methods, the integrals result in harmonic averages of a over each element, which can increase the accuracy of approximation.

In Chou and Li (1999), it has been shown for a covolume finite difference method for irregular meshes on convex domains that if the solution is in $H^2(\Omega)$, one can derive max-norm estimates $O(h^2 \ln h^{-1})$ for the error in the solution and $O(h)$ for the gradient (see also Li, Chen and Wu, 2000).

Earlier presentations of generalized finite difference methods can be found in Tikhonov and Samarskii (1962) and Heinrich (1987).

As shown in Kanschat and Rannacher (2002), asymptotic error expansions such as those used in Richardson extrapolation also hold for a uniform subdivision of a general coarse triangulation of a convex polygonal domain. The estimates for the remainder term are based on estimates for the discrete Green's function, similar to those found in Schatz and Wahlbin (1978).

7 A SUMMARY OF DIFFERENCE SCHEMES

For the convenience of the reader, we list below a summary of various difference approximations (Tables 1–4). For

Table 1. Basic finite difference approximations.

| Derivative | | Scheme | Accuracy |
|---------------------|-----------------------|--|---|
| u'_i | Forward | $\frac{1}{h}(u_{i+1} - u_i)$ | $O(h)$ |
| u'_i | Backward | $\frac{1}{h}(u_i - u_{i-1})$ | $O(h)$ |
| $b_1 u'_i$ | Upwind | if $b_1 > 0$, $\frac{b_1}{h}(u_i - u_{i-1})$, if $b_1 < 0$, $\frac{b_1}{h}(u_{i+1} - u_i)$ | $O(h)$ |
| u'_i | Central difference | $\frac{1}{2h}(u_{i+1} - u_{i-1})$ | $O(h^2)$ |
| u''_i | Forward | $\frac{1}{2h^2}(-3u_i + 4u_{i+1} - u_{i+2})$ | $O(h^2)$ |
| u''_i | Backward | $\frac{1}{2h^2}(3u_i - 4u_{i-1} + u_{i-2})$ | $O(h^2)$ |
| u''_i | Central difference | $\frac{1}{h^2}(u_{i+1} - 2u_i + u_{i-1})$ | $O(h^2)$ |
| u'''_i | Forward | $\frac{1}{h^3}(2u_i - 5u_{i+1} + 4u_{i+2} - u_{i+3})$ | $O(h^2)$ |
| u'''_i | Backward | $\frac{1}{h^3}(-u_{i-3} + 4u_{i-2} - 5u_{i-1} + 2u_i)$ | $O(h^2)$ |
| $u_{xx} + u_{yy}$ | Five-point cross | $\frac{1}{h^2}(u_{i-1,j} + u_{i+1,j} - 4u_{i,j} + u_{i,j+1} + u_{i,j-1})$ | $O(h^2)$ |
| $u_{xx} + u_{yy}$ | Five-point skewed | $\frac{1}{2h^2}(u_{i-1,j-1} + u_{i+1,j-1} - 4u_{i,j} + u_{i-1,j+1} + u_{i+1,j+1})$ | $O(h^2)$ |
| $u_{xx} + u_{yy}$ | Nine-point | $\frac{1}{6h^2}(4u_{i-1,j} + 4u_{i+1,j} - 20u_{i,j} + 4u_{i,j+1} + 4u_{i,j-1} + u_{i-1,j-1} - u_{i+1,j-1} + u_{i-1,j+1} + u_{i+1,j+1})$ | $O(h^4)$ |
| $au_{xx} + bu_{yy}$ | Nine-point anisotropy | $\frac{1}{6h^2}(\beta u_{i-1,j} + \beta u_{i+1,j} + \gamma u_{i,j+1} + \gamma u_{i,j-1} - 10(a+b)u_{i,j} + au_{i-1,j-1} + au_{i+1,j-1} + au_{i-1,j+1} + au_{i+1,j+1})$ | $O(h^4)$ if $\frac{1}{5}a \leq b \leq 5a$ $\alpha = \frac{a+b}{2}$ $\beta = 5\alpha - b$, $\gamma = 5b - a$ |

Table 2. Finite difference approximations for parabolic equations.

| Model parabolic problem: $u_t = bu_{xx} + f$ | | Accuracy/stability |
|--|--|---|
| Forward-time, central-space | $\frac{v_m^{k+1} - v_m^k}{k} = b \frac{v_m^k - 2v_m^{k-1} + v_m^{k-2}}{h^2} + f_m^k$ | Accuracy (k, h^2) ; stable if $\frac{bk}{h^2} \leq \frac{1}{2}$ |
| Backward-time, central-space | $\frac{v_m^{k+1} - v_m^k}{k} = b \frac{v_m^{k+1} - 2v_m^k + v_m^{k-1}}{h^2} + f_m^{k+1}$ | Accuracy (k, h^2) ; unconditionally stable |
| Crank–Nicolson | $\frac{v_m^{k+1} - v_m^k}{k} = \frac{1}{2}b \left[\frac{v_m^{k+1} - 2v_m^k + v_m^{k-1}}{h^2} + \frac{v_m^k - 2v_m^{k-1} + v_m^{k-2}}{h^2} \right] + \frac{1}{2}(f_m^{k+1} + f_m^k)$ | Accuracy (k^2, h^2) ; unconditionally stable (the θ -method, $\theta = 1/2$) |
| Leap-frog | $\frac{v_m^{k+1} - v_m^{k-1}}{2k} = b \frac{v_m^k - 2v_m^{k-1} + v_m^{k-2}}{h^2} + f_m^k$ | Unconditionally unstable |
| Du Fort–Frankel–Saul'ev | $\frac{v_m^{k+1} - v_m^{k-1}}{2k} = b \frac{v_m^{k+1} - (v_m^{k+1} + v_m^{k-1}) + v_m^{k-2}}{h^2} + f_m^k$ | Accuracy $v_{k,h} = O(h^2) + O(k^2) + O(h^{-2}k^2)$; explicit but unconditionally stable, consistent with a wrong problem, such as $u_t - u_{xx} + \alpha^2 u_{tt} = 0$ if $k = \alpha h$ and $\alpha > 0$ is fixed. |

Note: Accuracy (k^p, h^q) means that the local truncation error of the scheme is $\tau_{k,h} = O(k^p) + O(h^q)$.

Table 3. Finite difference approximations for first-order hyperbolic equations.

| Model hyperbolic problem: $u_t + au_x = f$ | | Accuracy/stability |
|--|---|--|
| Leap-frog | $\frac{v_m^{n+1} - v_m^{n-1}}{2k} + a \frac{v_m^n - v_m^{n-1}}{2h} = f_m^n$ | Conditionally stable $\left(\left \frac{a}{h} \right < 1 \right)$ |
| Lax–Friedrichs | $\frac{v_m^{n+1} - 1/2(v_m^{n+1} + v_m^{n-1})}{k} + a \frac{v_m^{n+1} - v_m^{n-1}}{2h} = f_m^n$ | Consistent for $k^{-1}h^2 \rightarrow 0$; CFL must hold (explicit scheme) $\left \frac{a}{h} \right \leq 1$ |
| Lax–Wendroff | $\frac{v_m^{n+1} - v_m^n}{k} + a \frac{v_m^{n+1} - v_m^{n-1}}{2h} - \frac{a^2 k}{2} \frac{v_m^{n+1} - 2v_m^n + v_m^{n-1}}{h^2} = f_m^n$ $= \frac{1}{2}(f_m^{n+1} + f_m^n) - \frac{ak}{4h}(f_m^{n+1} - f_m^{n-1})$ | Accuracy (k^2, h^2) ; CFL $\left \frac{a}{h} \right \leq 1$ |
| Crank–Nicolson | $\frac{v_m^{n+1} - v_m^n}{k} + a \frac{v_m^{n+1} - v_m^{n-1} + v_m^n - v_m^{n-1}}{4h} = \frac{f_m^{n+1} + f_m^n}{2}$ | Accuracy (k^2, h^2) ; implicit scheme (uncond. stable) |
| Forward–backward MacCormack | $\tilde{v}_m^{n+1} = v_m^n - a \frac{k}{h^2}(v_{m+1}^n - v_m^n) + kf_m^n$ $v_m^{n+1} = \frac{1}{2} \left[v_m^n + \tilde{v}_m^{n+1} - a \frac{k}{h^2}(\tilde{v}_m^{n+1} - \tilde{v}_{m-1}^{n+1}) + kf_m^{n+1} \right]$ | Accuracy (k^2, h^2) ; identical to Lax–Wendroff when f is constant |

Note: Accuracy (k^p, h^q) means that the local truncation error of the scheme is $\tau_{k,h} = O(k^p) + O(h^q)$.

Table 4. Finite difference approximations for second-order hyperbolic equations.

| Model hyperbolic problem: $u_{tt} = a^2 u_{xx} + f$ | |
|---|--|
| Scheme | Accuracy/stability |
| The γ -method, (2, 2) accurate | |
| $(1 - \gamma c^2 k^2 \Delta_t)(v_m^{n+1} - 2v_m^n + v_m^{n-1}) = c^2 k^2 \Delta_t v_m^n + k^2 [\gamma f_m^{n+1} + (1 - 2\gamma) f_m^n + \gamma f_m^{n-1}]$ | Unconditionally stable for $\gamma \geq 1/4$ Conditionally stable for $0 \leq \gamma < 1/4$ |
| 1st (2, 4) accurate scheme (for $f = 0$) | |
| $\frac{v_m^{n+1} - 2v_m^n + v_m^{n-1}}{k^2} = a^2 \left(1 - \frac{h^2}{12} \Delta_x^2 \right) v_m^n$ $= a^2 \left[-\frac{v_{m+2}^n + 16v_{m+1}^n - 30v_m^n + 16v_{m-1}^n - v_{m-2}^n}{12h^2} \right]$ | Conditionally stable: $\frac{k}{h} \leq \frac{\sqrt{3}}{2}$ |
| 2nd (2, 4) accurate scheme (for $f = 0$) | |
| $\Delta_t^2 v_m^n = a^2 \left(1 + \frac{h^2}{12} \Delta_x^2 \right) \Delta_t^2 v_m^n$, or $v_{m+1}^{n+1} + 10v_m^{n+1} + v_{m-1}^{n+1} - 2(v_{m+1}^n + v_{m-1}^n) - 10v_m^n + v_{m-1}^n$ $+ v_{m+1}^{n-1} + 10v_m^{n-1} + v_{m-1}^{n-1} = 12a^2 \frac{k^2}{h^2} (v_{m+1}^n - 2v_m^n + v_{m-1}^n)$ | Conditionally stable: $\frac{k}{h} \leq \sqrt{\frac{2}{3}}$ |

Note: Accuracy (p, q) means that the local truncation error of the scheme is $\tau_{k,h} = O(k^p) + O(h^q)$.
Symbolic notation: $\Delta_t^2 = [(v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1})/h^2]$.

a derivation and analysis of these schemes, see Richtmeyer and Morton (1967) and Strikwerda (1989).

REFERENCES

- Axelsson O. Global integration of differential equations through Lobatto quadrature. *BIT* 1964; 4:69–86.
- Axelsson O. Error estimates over infinite intervals of some discretizations of evolution equations. *BIT* 1984; 24:413–424.
- Axelsson O. *Iterative Solution Methods*. Cambridge University Press: New York, 1994.
- Axelsson O and Barker VA. *Finite Element Solution of Boundary Value Problems. Theory and Computation*, Computer Science and Applied Mathematics. Academic Press, 1984.
- Axelsson O and Gustafsson I. An iterative solver for a mixed variable variational formulation of the (first) biharmonic problem. *Comput. Methods Appl. Mech. Eng.* 1979; 20:9–16.
- Axelsson O and Gustafsson I. An efficient finite element method for nonlinear diffusion problems. *Bull. Greek Math. Soc.* 1991; 32:45–61.
- Axelsson O and Kolotillina L. Monotonicity and discretization error estimates. *SIAM J. Numer. Anal.* 1990; 27:1591–1611.
- Axelsson O and Kucherov A. Real valued iterative methods for solving complex symmetric linear systems. *Numer. Lin. Alg. Appl.* 2000; 7:197–218.
- Axelsson O and Layton W. Defect-correction for convection dominated convection–diffusion problems. *Math. Modell. Numer. Anal., M2AN* 1990; 24:423–455.
- Axelsson O and Marinova R. A hybrid method of characteristics and central difference method for convection–diffusion problems. *Comput. Appl. Math.* 2002; 21:631–659.
- Axelsson O and Nikolova M. Avoiding slave points in an adaptive refinement procedure for convection–diffusion problems in 2D. *Computing* 1998; 61:331–357.
- Axelsson O and Vervier JG. Boundary value techniques for initial value problems in ordinary differential equations. *Math. Comp.* 1984; 45:153–171.
- Bakhvalov AN. On the optimization of methods for solving boundary value problems with boundary layers. *Zh. Vychisl. Math. i Mat. Fiz.* 1969; 9:841–859.
- Benson RC and Bogy DB. Deflection of a very flexible spinning disk due to a stationary transverse load. *J. Appl. Mech., Trans. ASME* 1978; 45:636–642.
- Brugnano L and Trigiante D. *Solving Differential Problems by Multistep Initial and Boundary Value Methods*. Gordon & Breach Science Publishers, 1998.
- Blum H, Lin Q and Rannacher R. Asymptotic error expansion and Richardson extrapolation for linear finite elements. *Numer. Math.* 1986; 49:11–37.
- Brezzi F and Fortin M. *Mixed and Hybrid Finite Element Methods*. Springer: New York, 1991.
- Butcher JC. On A-stable implicit Runge-Kutta methods. *BIT* 1977; 17:375–378.
- Chou S-H and Li Q. Error estimates in L₂, H¹ and L (in covolume) methods for elliptic and parabolic problems: a unified approach. *Math. Comp.* 1999; 69:103–120.
- Ciarlet PG and Raviart PA. A mixed finite element method for the biharmonic equation. In *Proceedings of a Symposium on Mathematical Aspects of Finite Elements in Partial Differential Equations*, de Boor C (ed.). Academic Press: New York, 1974; 125–145.
- Collatz L. Einige Anwendungen der mehrdimensionalen Approximationstheorie zur Lösungseinschließung bei Randwertaufgaben. *Differential Equations and their Applications*, Equadiff 6; Proceedings of the 6th International Conference, Bmo, 1985; *Lecture Notes Math.* Springer-Verlag: Berlin, Heidelberg, New York, 1986; 1192:367–372.
- Courant R, Friedrichs KO and Lewy H. Über die Partiiellen Differenzengleichungen der Mathematischen Physik. *Math. Ann.* 1928; 100:31–74.
- Dahlquist G. A special stability problem for linear multistep methods. *BIT* 1963; 3:27–43.
- Dahlquist G. On accuracy and unconditional stability of linear multistep methods for second order differential equations. *BIT* 1978; 18:133–136.
- Douglas J Jr. and Russell TF. Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures. *SIAM J. Numer. Anal.* 1982; 19:871–885.
- Duarte CA and Oden JT. H-p clouds and h-p meshless methods. *Numer. Methods Part. Differ. Equations* 1996; 20:63–71.
- Garanzha V. Variational principles in grid generation and geometric modelling: theoretical justifications and open problems. *Numer. Lin. Alg. Appl.* 2004; 11: (to appear).
- George A and Lin JWH. *Computer Solution of Large Sparse Positive Definite Systems*. Prentice Hall, 1981.
- Gustafson K and Abe T. The third boundary condition – was it Robin's? *The Math. Intell.* 1998; 20:63–71.
- Hackbusch W. *Multigrid Methods and Applications*, Springer Series in Computational Mathematics, vol. 4. Springer-Verlag: Berlin, 1985.
- Hairer E. Unconditionally stable methods for second order differential equations. *Numer. Math.* 1979; 32:373–379.
- Heinrich B. *Finite Difference Methods for Irregular Networks: A Generalized Approach to Second Order Elliptic Problems*, International Series of Numerical Mathematics, vol. 82. Birkhäuser, 1987.
- Houstis EN and Rice JR. High order methods for elliptic partial differential equations with singularities. *Int. J. Numer. Methods Eng.* 1982; 18:737–754.
- Iserles A. *Numerical Analysis of Differential Equations*. Cambridge University Press, 1996.
- Kanschat G and Rannacher R. Local error analysis of the interior penalty discontinuous Galerkin method for second order elliptic problems. *J. Numer. Math.* 2002; 10:249–274.
- Linus T and Synes M. A hybrid difference scheme on a Shishkin mesh for linear convection–diffusion problems. *Appl. Numer. Math.* 1999; 31:255–270.
- Lamb H and Southwell RV. The vibrations of a spinning disk. *Proc. R. Soc., London* 1921; 99:272–280.
- Li R, Chen Zh and Wu W. *Generalized Difference Methods for Differential Equations*. Marcel Dekker, 2000.
- Peaceman DW and Rachford HH Jr. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Ind. Appl. Math.* 1955; 3:28–41.
- Richtmeyer R and Morton K. *Difference Methods for Initial Value Problems*. Interscience Publishers: New York, 1967.
- Schatz AH and Wahlbin LB. Maximum norm estimates in the finite element method on plane polygonal domains. I. *Math. Comp.* 1978; 32:73–109.
- Shih SD and Kellogg RB. Asymptotic analysis of a singularly perturbation problem. *SIAM J. Math. Anal.* 1987; 18: 567–579.
- Shishkin G. *Grid Approximation of Singularly Perturbed Elliptic and Parabolic Equations*, Second Doctoral Thesis. Keldysh Institute, Russian Academy of Science: Moscow, 1990 (in Russian).
- Shortley GH and Weller R. The numerical solution of Laplace's equation. *J. Appl. Phys.* 1938; 9:334–348.
- Strikwerda JC. *Finite Difference Schemes and Partial Differential Equations*. Chapman & Hall, 1989.
- Tikhonov AN and Samarskii AA. Homogeneous difference schemes on nonuniform nets. *Zh. Vychisl. Math. i Mat. Fiz.* 1962; 2:812–832. English translation in USSR: *Comput. Math. Math. Phys.* 1962; 2:927–953.
- Thompson JP, Warsi ZUA and Mastin CW. *Numerical Grid Generation – Foundations and Applications*. North Holland, 1985.
- Varga R. *Matrix Iterative Analysis*. Prentice Hall: Englewood Cliffs, 1962.
- Yanenko NN. In *The Method of Fractional Steps*. English translation Holt M (ed.). Springer-Verlag: New York, 1971.

FURTHER READING

- Forsythe GE and Wasow WR. *Finite-Difference Methods for Partial Differential Equations*, Applied Mathematics Series. John Wiley & Sons: New York, London, 1960.
- Fried I. On a deficiency in unconditionally stable explicit time-integration methods in elastodynamics and heat transfer. *Comput. Methods Appl. Mech. Eng.* 1984; 46:195–200.
- Gustafsson B, Kreiss H-O and Olliger J. *Time Dependent Problems and Finite Difference Methods*. John Wiley & Sons: New York, 1995.
- Henrici P. *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley: New York, 1962.
- Keller HB. *Numerical Methods for Two-Point Boundary Value Problems*. Blaisdell: London, 1968.
- Lax PD and Wendroff B. Difference schemes for hyperbolic equations with high order of accuracy. *Commun. Pure Appl. Math.* 1964; 17:381–391.
- Meiss T and Marcowitz U. *Numerische Behandlung Partieller Differentialgleichungen*. Springer-Verlag: Berlin, 1978.
- Morton KW. *Numerical Solution of Convection–Diffusion Problems*. Chapman & Hall: London, 1996.

Strang G. *Linear Algebra*. John Wiley: New York, 1976.

Roes H, Symes M and Tobiska L. *Numerical Methods for Singularly Perturbed Differential Equations*. Springer: Heidelberg, 1996.

Trefethen LN. Group velocity in finite difference schemes. *SIAM Rev.* 1982; 24:113–136.

Chapter 3

Interpolation in h -version Finite Element Spaces

Thomas Apel

Universität der Bundeswehr München, Neubiberg, Germany

| | |
|---|----|
| 1 Introduction | 55 |
| 2 Finite Elements | 56 |
| 3 Definition of Interpolation Operators | 58 |
| 4 The Deny–Lions Lemma | 60 |
| 5 Local Error Estimates for the Nodal Interpolant | 61 |
| 6 Local Error Estimates for Quasi-Interpolants | 66 |
| 7 Example for a Global Interpolation Error Estimate | 68 |
| 8 Related Chapters | 70 |
| References | 70 |

1 INTRODUCTION

The aim of this chapter is to discuss interpolation operators that associate with a function u , an element from an h -version finite element space. We investigate nodal interpolation and several variants of quasi-interpolation and estimate the interpolation error. The discussion becomes diverse because we include (to a different extent) triangular/tetrahedral and quadrilateral/hexahedral meshes, affine and nonaffine elements, isotropic and anisotropic elements, and Lagrangian and other elements.

Interpolation error estimates are used for a priori and a posteriori estimation of the discretization error of a finite element method. This is explained in other chapters

of the Encyclopedia (see Chapter 4, this Volume for finite element methods in the displacement formulation or Chapter 9, this Volume for mixed finite element methods). To estimate the error a priori, one can often use the *nodal interpolant*. To get optimal error bounds, one has to use the maximum available regularity of the solution. Since the regularity can be described differently, one is interested in local interpolation error estimates with various norms on the right-hand side, including norms not only in the classical Sobolev spaces but also in weighted Sobolev spaces or in Sobolev–Slobodetskii spaces. For the ease of presentation, this chapter is restricted to the case of Sobolev spaces.

The situation is different for a posteriori error estimates. In order to investigate residual-type error estimators, local interpolation error estimates for functions from the Sobolev space $W^{1,2}(\Omega)$ are needed. Such estimates can be obtained for many finite elements only for *quasi-interpolation* operators where point values of functions or derivatives are replaced in the definition of the interpolation operators by certain averaged values.

In Section 2, we introduce finite elements and meshes. Section 3 is devoted to the definition of the interpolation operators. After a short discussion of the classical Deny–Lions lemma in Section 4, we derive error estimates for the nodal interpolation operator in Section 5, both for isotropic and anisotropic elements. We develop the theory in detail for affine elements and discuss shortly the non-affine case. Quasi-interpolants are investigated in Section 6 for isotropic Lagrangian elements, whereas anisotropic elements are mentioned only in brief. An example for a global interpolation error estimate is presented in Section 7. A typical solution with corner singularities is interpolated on a family of graded meshes, which is chosen such that the optimal order of convergence is obtained despite the irregular terms.

Encyclopedia of Computational Mechanics, Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes. Volume 1: *Fundamentals*. © 2004 John Wiley & Sons, Ltd. ISBN: 0-470-84699-2.

We refer to the literature at the appropriate places in this overview chapter and omit references to related work in the introduction. Moreover, we underline that the chapter is written in the spirit of the h -version of the finite element method. We do not investigate the dependence of constants on the polynomial degree of the functions. For interpolation error estimates in the context of the p - and hp -versions of the finite element method, we refer, for example, to Schwab (1998) for an overview and to Melenk (2003) for more recent work on quasi-interpolation.

Let $d = 2, 3$ be the space dimension and $x = (x_1, \dots, x_d)$ a Cartesian coordinate system. We use standard multi-index notation with $\alpha := (\alpha_1, \dots, \alpha_d)$, where the entries α_i are from the set \mathbb{N}_0 of nonnegative integers, and

$$x^\alpha := \prod_{i=1}^d x_i^{\alpha_i}, \quad \alpha! := \prod_{i=1}^d \alpha_i!, \quad |\alpha| := \sum_{i=1}^d \alpha_i$$

$$D^\alpha := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}$$

The notation $W^{\ell,p}(G)$, $\ell \in \mathbb{N}_0$, $p \in [1, \infty]$, is used for the classical Sobolev spaces with the norm and seminorm

$$\|v\|_{W^{\ell,p}(G)} := \sum_{|\alpha| \leq \ell} \int_G |D^\alpha v|^p$$

$$|v|_{W^{\ell,p}(G)} := \sum_{|\alpha| = \ell} \int_G |D^\alpha v|^p$$

for $p < \infty$ and the usual modification for $p = \infty$. In general, we will write $L^p(G)$ for $W^{0,p}(G)$. The symbol C is used for a generic positive constant, which may be of a different value at each occurrence. C is always independent of the size of finite elements, but it may depend on the polynomial degree.

2 FINITE ELEMENTS

In this section, we introduce in brief the notion of the finite element. While Chapter 4, this Volume presents this topic comprehensively, we focus on those pieces that are necessary for the remaining part of the current chapter.

Charlet (1978) introduces the finite element as the triple $(K, \mathcal{P}_K, \mathcal{N}_K)$, where K is a closed bounded subset of \mathbb{R}^d (some authors use open sets) with nonempty interior and piecewise smooth boundary, \mathcal{P}_K is an n -dimensional linear space of functions defined on K , and $\mathcal{N}_K = \{N_{i,K}\}_{i=1}^n$ is a basis of the dual space \mathcal{P}_K^* . The smooth parts of the boundary are called *faces*; they meet in *edges* and *vertices*. In two dimensions, the edges play the role of the faces.

The functions in \mathcal{P}_K and the functionals in \mathcal{N}_K are sometimes called *shape functions* and *nodal variables*

respectively. We adopt these names here, although \mathcal{P}_K does not necessarily define the shape of the element K (this is true only for *isoparametric elements*; see below) and also $N \in \mathcal{N}_K$ is not necessarily a function evaluation in nodes. Important examples of finite elements are discussed in Chapter 4, this Volume (Examples 4–9 and 16–17). The nodal variables define a basis $\{\phi_{i,K}\}_{i=1}^n$ of \mathcal{P}_K via

$$N_{i,K}(\phi_{j,K}) = \delta_{i,j}, \quad i, j = 1, \dots, n \quad (1)$$

which is called the *nodal basis*. This basis is employed here since it allows an elegant definition of interpolation operators. Note, however, that other bases might be advantageous for a hierarchical approach to approximation (see e.g. Chapter 7, this Volume) or for an efficient implementation (see e.g. Chapter 5, this Volume).

Example 1. An important example is the element $(K, \mathcal{P}_K, \mathcal{N}_K)$ with a triangular or tetrahedral set K , with $\mathcal{P}_K = \mathcal{P}_1$ being the space of polynomials of degree one, and with the set \mathcal{N}_K consisting of function evaluations at the vertices of K . The nodal basis can be represented here by the barycentric coordinates $\lambda_{j,K}$ of the element, $\phi_{j,K} = \lambda_{j,K}$, $j = 1, \dots, d+1$. Recall that any point $x \in K$ defines (together with the vertices of K) a splitting of the element into $d+1$ triangles/tetrahedra $K_j(x)$. The ratios $\lambda_{j,K} := |K_j(x)|/|K|$ are called barycentric coordinates of x .

Finite element spaces over a domain $\Omega \subset \mathbb{R}^d$ are defined on the basis of a *finite element mesh* or *triangulation* \mathcal{T} of Ω . This is a subdivision of Ω into a finite number of closed bounded subsets K with nonempty interior and piecewise smooth boundary in such a way that the following properties are satisfied:

1. $\bar{\Omega} = \bigcup_{K \in \mathcal{T}} K$.
2. Distinct $K_1, K_2 \in \mathcal{T}$ have no common interior points.
3. Any face of an element $K_1 \in \mathcal{T}$ is either a subset of the boundary $\partial\Omega$ or a face of another element $K_2 \in \mathcal{T}$.

Let any element K be associated with a finite element $(K, \mathcal{P}_K, \mathcal{N}_K)$. We define the corresponding finite element space by

$$\text{FE}_{\mathcal{T}} := \left\{ v \in L^2(\Omega) : \begin{array}{l} v_K := v|_K \in \mathcal{P}_K \quad \forall K \in \mathcal{T} \text{ and} \\ v_{K_1}, v_{K_2} \text{ share the same nodal} \\ \text{values on } K_1 \cap K_2 \end{array} \right\} \quad (2)$$

(compare Chapter 4, this Volume). The dimension of $\text{FE}_{\mathcal{T}}$ is denoted by $N_{\mathcal{T}}$.

Remark 1. There are other approaches to the construction of finite element discretizations, for example, the *weighted*

extended B-spline approximation where the geometry is not described by the mesh but by weight functions (see Höllig (2003) for an overview). Since this method does not fall naturally into the frame we are developing here, we will not discuss this method in detail.

The weak solution of elliptic boundary value problems of order $2m$ is generally searched in a subspace V of the Sobolev space $W^{m,2}(\Omega)$. The space V is defined by imposing appropriate (for simplicity, homogeneous) boundary conditions. By imposing these boundary conditions also in the finite element space $\text{FE}_{\mathcal{T}}$, we obtain the N -dimensional finite element space $V_{\mathcal{T}}$. A *conforming* finite element method requires $V_{\mathcal{T}} \subset V$, a condition that is satisfied only when the nodal values imply certain smoothness of the finite element functions, in particular, $\text{FE}_{\mathcal{T}} \subset C^{m-1}(\Omega)$ (see also Chapter 4, this Volume). On the other hand, solutions to electromagnetic field problems are not necessarily in $W^{1,2}(\Omega)$ such that the approximating spaces also need not be continuous (see the survey of Hipmair, 2002). Continuous finite element spaces might even lead to wrong solutions in this case.

For later use, we also define in $\text{FE}_{\mathcal{T}}$ and $V_{\mathcal{T}}$ the global sets of nodal variables (functionals) $\mathcal{N}_{\mathcal{T},+} = \{N_{i,K}\}_{i=1}^n$ and $\mathcal{N}_{\mathcal{T}} = \{N_{i,K}\}_{i=1}^n$ respectively, and the corresponding global basis $\{\phi_{i,K}\}_{i=1}^n \subset \text{FE}_{\mathcal{T}}$ that satisfies

$$N_i(\phi_j) = \delta_{i,j}, \quad i, j = 1, \dots, N_{\mathcal{T}}$$

The set $\mathcal{N}_{\mathcal{T},+}$ is the union of all \mathcal{N}_K , $K \in \mathcal{T}$, where common nodal variables of adjacent elements are counted only once. In $\mathcal{N}_{\mathcal{T},+}$, some nodal variables (degrees of freedom) might be suppressed because of the boundary conditions. Note further that $V_{\mathcal{T}}$ is spanned by $\{\phi_{i,K}\}_{i=1}^n$.

In this chapter, we will concentrate on triangulations consisting of simplices (triangles or tetrahedra) or convex quadrilaterals/hexahedra. A typical parameter for the description of the elements K is the *aspect ratio* γ_K , the ratio of the diameter h_K of K , and the diameter q_K of the largest ball inscribed in K . We will call elements with a moderate aspect ratio *isotropic* and elements with large aspect ratio *anisotropic*. For isotropic elements, we allow the quantity γ_K to be included in constants in error estimates, whereas for anisotropic elements, the aspect ratio must be separated from constants, which means constants must be uniformly bounded in the aspect ratio.

Example 2 (Isotropic and anisotropic simplices) Triangles and tetrahedra with plane faces (edges) are sometimes called *shape-regular* if they are isotropic. Shape-regularity is generally used as a property that is easy to achieve in mesh generation and that allows for a numerical analysis (e.g. interpolation error estimation and also the proof of

a discrete inf-sup condition in mixed methods or efficient multilevel solvers) at moderate technical expense.

Zlámal (1968) has shown for triangles with straight edges that a lower bound on the interior angles is equivalent to an upper bound on the aspect ratio. Therefore, shape-regularity can be defined equivalently via a *minimal angle condition*: There exists a constant $\gamma_{\min} > 0$ such that the angles of all triangles of a family of triangulations are bounded from below by γ_{\min} .

Elements with large aspect ratio can be used advantageously for the approximation of anisotropic features in functions (solutions of boundary value problems), for example, boundary layers or singularities in the neighborhood of concave edges of the domain. For the numerical analysis, it is often necessary to impose a *maximal angle condition*: There exists a constant $\gamma_{\max} > 0$ such that the angles of all triangles of a family of triangulations are bounded from above by γ_{\max} . An analogous definition can be given for tetrahedra (see Apel, 1999a, pages 54, 90f). Figure 1 shows an isotropic triangle and two anisotropic triangles, one that satisfies the maximal angle condition and one that does not. Note that if the angles are bounded from below away from zero, they are also bounded from above away from zero, whereas the converse is not true. Therefore, estimates are usually easier to obtain for shape-regular elements (where $\cot \gamma_{\min}$ enters the constant) than for anisotropic elements (where, if necessary at all, $\cot \gamma_{\max}$ enters the constants).

Most monographs consider only shape-regular elements, for example, Braess (1997), Brenner and Scott (1994), Ciarlet (1978, 1991), Hughes (1987), and Oswald (1994). Anisotropic elements are investigated mainly in research papers and in the book by the author Apel (1999a). The maximal angle condition was introduced first by Sygne (1957), and later rediscovered by Gregory (1975), Babuska and Aziz (1976), and Jamet (1976).

Example 3 (Shape-regular quadrilaterals) From the theoretical point of view, it is important to distinguish between parallelograms and more general quadrilaterals. Parallelograms share the following property with triangles: for two elements K_1 and K_2 , there is an invertible affine mapping $F: \hat{x} \in \mathbb{R}^d \mapsto x = F(\hat{x}) = A\hat{x} + a \in \mathbb{R}^d$ with $K_2 = F(K_1)$. This property simplifies proofs, and results for triangles can usually be extended to parallelo-



Figure 1. Isotropic and anisotropic triangles.



Figure 2. Degenerated isotropic quadrilaterals.

grams. Shape-regularity is defined by a bounded aspect ratio γ_K . Parallelograms are sometimes even easier to handle than triangles since the edges point into two directions only. Similar statements can be made in the three-dimensional case for parallelepipeds.

The situation changes for more general elements. A bounded aspect ratio is necessary but not sufficient for shape-regular quadrilaterals. For several estimates, it is advantageous to exclude quadrilaterals that degenerate to triangles (see Figure 2). The literature is not unanimous about an appropriate description. Ciarlet and Raviart (1972a,b) demand a uniformly bounded ratio of the lengths of the longest and the shortest edge of the quadrilateral and that the interior angles are away from zero and π . Girault and Raviart (1986) assume equivalently that the four triangles that can be formed from the vertices of the quadrilateral are shape-regular in the sense of Example 2. Another equivalent (more technical) version is given by Arunakirinathar and Reddy (1995).

Weaker mesh conditions were derived by Jamet (1977) and Acosta and Durán (2000). Jamet proves that the elements shown in Figure 2 can be admitted, but he still relies on a bounded aspect ratio. Acosta and Durán formulate the *regular decomposition property*, which is the weakest known condition that allows to prove the standard interpolation error estimate for Q_1 elements. For a detailed review, we refer to Ming and Shi (2002a,b).

Further classes of meshes can be described as being asymptotical parallelograms (see also the papers by Ming and Shi). Some results that are valid for parallelograms can be extended to such meshes but not to general quadrilateral meshes, for example, superconvergence results and interpolation error estimates for certain serendipity elements (compare Remark 7). Meshes of one of these classes arise typically from a hierarchical refinement of a coarse initial mesh.

A more detailed discussion of all these conditions is beyond the frame of this chapter. We will restrict further discussion to affine elements and to elements that are shape-regular in the sense of Ciarlet/Raviart or Girault/Raviart.

We will develop the interpolation theory on the basis of the following assumption.

Assumption 1. The finite element space FE_T is constructed on the basis of a reference element $(\hat{K}, \mathcal{P}_{\hat{K}}, \mathcal{N}_{\hat{K}})$ by the following rule:

1. For each $K \in \mathcal{T}$, there is a bijective mapping $F_K: \hat{x} \in \mathbb{R}^d \mapsto x = F_K(\hat{x}) \in \mathbb{R}^d$ with $K = F_K(\hat{K})$,
2. $u \in \mathcal{P}_K$ iff $\hat{u} := u \circ F_K \in \mathcal{P}_{\hat{K}}$, and
3. $N_{i,K}(u) = N_{i,\hat{K}}(u \circ F_K)$, $i = 1, \dots, n$, for all u for which the functionals are well defined.

If possible, an affine mapping is chosen,

$$F_K(\hat{x}) = A\hat{x} + a \quad \text{with } A \in \mathbb{R}^{d \times d}, a \in \mathbb{R}^d$$

otherwise the *isoparametric mapping* is used,

$$F_K(\hat{x}) = \sum_{i=1}^n a^i \psi_{i,\hat{K}}(\hat{x})$$

where the shape of K is determined by the positions of nodes $a^i \in \mathbb{R}^d$ and shape functions $\psi_{i,\hat{K}}$ with $\psi_{i,\hat{K}}(a^j) = \delta_{i,j}$. A typical example (in particular for Lagrangian elements, these are elements with $N_i(u) = u(a^i)$ where a^i are the nodes of K) is to choose $\text{span}\{\psi_{i,\hat{K}}\}_{i=1}^n = \mathcal{P}_{\hat{K}}$ but one can also choose a lower-dimensional space. Then the mapping is called *subparametric*. An example is to use Q_1 instead of Q_2 in the case of quadrilateral or hexahedral elements.

These examples for the mapping show that Assumption 1 is not too restrictive. Affine and isoparametric Lagrangian elements are covered. Note, in particular, that the space \mathcal{P}_K for nonsimplicial elements (quadrilaterals, hexahedra, ...) or for isoparametric elements is generally defined via (2). For non-Lagrangian elements, condition (3) is the critical one. If \mathcal{N}_K contains the evaluation of the gradient in a vertex of K , the nodal variable should be written in the form of scaled derivatives in the directions of the edges that meet in this vertex; see also the discussion in Brenner and Scott (1994), Section 3.4. Analogously, functionals in the form of integrals should be properly scaled. Such technical manipulation, however, is not possible in the case of normal derivatives; they are transformed into oblique derivatives. These elements are excluded by our framework but they can sometimes be treated as a perturbation of another element that is conforming with the theory we are going to present; see, for example, the estimates for the Argyris element in Ciarlet (1978, pages 337ff).

3 DEFINITION OF INTERPOLATION OPERATORS

3.1 Nodal interpolation

Given a finite element $(K, \mathcal{P}_K, \mathcal{N}_K)$ with a nodal basis $\{\phi_{i,K}\}_{i=1}^n$ and the nodal variables $\{N_{i,K}\}_{i=1}^n$, it is straightforward to introduce the nodal interpolation operator

$$I_K u := \sum_{i=1}^n N_{i,K}(u) \phi_{i,K}$$

The duality condition (1) yields $I_K \phi_{j,K} = \phi_{j,K}$, $j = 1, \dots, n$, and thus

$$I_K \phi = \phi \quad \forall \phi \in \mathcal{P}_K \quad (3)$$

Under Assumption 1, we obtain the property

$$\begin{aligned} (I_K u) \circ F_K &= \left(\sum_{i=1}^n N_{i,K}(u) \phi_{i,K} \right) \circ F_K \\ &= \sum_{i=1}^n N_{i,\hat{K}}(\hat{u}) \phi_{i,\hat{K}} = I_{\hat{K}} \hat{u} \end{aligned}$$

that allows the estimation of the error on the reference element \hat{K} and the transformation of the estimates to K . The interpolation operator I_K is well defined for functions u that allow the evaluation of the functionals $N_{i,K}$. For example, if these functionals include the pointwise evaluation of derivatives up to order s , then I_K is defined for functions from $C^s(K) \subset W^{s,p}(K)$ with $s^* > s + d/p$. If the functionals include the evaluation of integrals only, the required smoothness is correspondingly lower.

The definition of FE_T in (2) allows the introduction of the global interpolation operator I_T by

$$(I_T u)|_K = I_K(u|_K) \quad \forall K \in \mathcal{T}$$

With the basis $\{\phi_{i,K}\}_{i=1}^{N_K}$ of FE_T and the globalized set of nodal variables $\mathcal{N}_{T,K} = \{N_{i,K}\}_{i=1}^{N_K}$, we can write

$$I_T u = \sum_{i=1}^{N_K} N_{i,K}(u) \phi_{i,K} \quad (4)$$

Note again that I_T acts only on sufficiently regular functions such that all functionals N_i are well defined.

Remark 2. We distinguish between the finite element space FE_T of dimension N_K and its N -dimensional subspace V_T where boundary conditions are imposed, $N_i(u) = 0$ for $u \in V$ and $i = N+1, \dots, N_K$. Therefore, equation (4) is equivalent to

$$I_T u = \sum_{i=1}^N N_i(u) \phi_i \quad (5)$$

that means boundary conditions imply no particular difficulty for the analysis of the nodal interpolation operator.

Remark 3. There are also interpolation operators that cannot be treated in the framework developed in this section; for example, Demkowicz's projection-based interpolant (see Oden, Demkowicz, Westermann, and Rachowicz, 1989), which is not defined by nodal variables but by a $(d+1)$ -step procedure with $W_0^{1,2}$ -projections on edges, faces, and volumes.

3.2 Quasi-interpolation

A drawback of nodal interpolation is the required regularity of the functions the operator acts on. For example, for Lagrangian elements, we need $u \in W^{s,p}(K)$ with $s^* > d/p$ to obtain well-defined point values via the Sobolev embedding theorem. This assumption may fail even for simple problems like the Poisson problem with mixed boundary conditions in concave three-dimensional domains, where a r^* -singularity with λ close to 0.25 may occur. Moreover, an interpolation operator for $W^{1,2}(\Omega)$ -functions is needed for the analysis of a posteriori error estimators and multilevel methods.

The remedy is the definition of a quasi-interpolation operator

$$Q_T u = \sum_{i=1}^N N_i(\Pi_i u) \phi_i \quad (6)$$

that means we replace the function u in (5) by the regularized functions $\Pi_i u$. The index i indicates that we may use for each functional N_i a different, locally defined averaging operator Π_i .

For simplicity of the exposition, we restrict ourselves to Lagrangian finite elements, that is, the nodal variables have the form $N_i(u) = u(a^i)$, where a^i are nodes in the mesh. For quasi-interpolation of C^1 -elements, we refer to Girault and Scott (2002), and for the definition of quasi-interpolants for lowest-order Nédélec elements of first type and lowest-order Raviart–Thomas elements that fulfill the *commuting diagram property* (de Rham diagram), we refer to Schöberl (2001).

Each node a^i , $i = 1, \dots, N$, is now related to a subdomain $\omega_i \subset \Omega$ and a finite-dimensional space \mathcal{P}_i . Different authors prefer different choices. We present two main examples.

Example 4 (Clément operator) Clément (1975) considers $\mathcal{P}_K = \mathbb{P}_2$ in simplicial elements K with plane faces. Each node a^i , $i = 1, \dots, N$, is related to the subdomain $\omega_i := \text{int supp } \phi_i$, where ϕ_i is the corresponding nodal basis function and *int* stands for interior. The averaging operator

$$\Pi_i: L^1(\omega_i) \rightarrow \mathbb{P}_{L-1} \quad (7)$$

is then defined by

$$\int_{\omega_i} (v - \Pi_i v) \phi = 0 \quad \forall \phi \in \mathbb{P}_{\ell-1} \quad (8)$$

which is for $v \in L^2(\omega_i)$ the $L^2(\omega_i)$ -projection into $\mathbb{P}_{\ell-1}$. This operator has the important property

$$\Pi_i \phi = \phi \quad \forall \phi \in \mathbb{P}_{\ell-1} \quad (9)$$

One can choose the parameter ℓ in correspondence with k , for example, $\ell = k + 1$, or in correspondence with the regularity of u . For $u \in W^{s,p}(\Omega)$, the choice $\ell = \min\{s, k + 1\}$ is appropriate.

We analyze interpolation error estimates for the resulting operator Q_T (see (6) in Section 6). Note that $Q_T v|_K$ is not only determined by $v|_K$ but also by $v|_{\omega_K}$.

$$\omega_K := \bigcup_{i: \omega_i \in K} \omega_i \quad (10)$$

Remark 4. There are several modifications of the Clément operator from Example 4. Bernardi (1989) computes the average in some reference domain $\hat{\omega}_i$, which is chosen from a fixed set of reference domains. This idea is used to treat meshes with curved (isoparametric) simplicial and quadrilateral elements. The particular difficulty is that the transformation that maps $\hat{\omega}_i$ to ω_i is only piecewise smooth. Bernardi and Girault (1998) and Carstensen (1999) modify further and project into spaces of piecewise polynomial functions.

A particularly simple choice $\omega_i = \text{int } K_i$ is used by Oswald (1994), who uses just one element $K_i \in \mathcal{T}$ with $a^i \in K_i$. In this way, the technical difficulty mentioned above is avoided.

Verfürth (1999b) develops the new projection operator $P_G^{\ell-1}$ (see the last paragraph in Section 4) and uses it in Verfürth (1999a) as the averaging operator Π_i in the definition of a quasi-interpolation operator. This modification allows for making explicit the constants in the error estimates.

Remark 5. The quasi-interpolant $Q_T v$ satisfies Dirichlet boundary conditions by construction since the sum in (6) extends only over the N degrees of freedom of V_T . The nice property of I_T mentioned in Remark 2 is not satisfied for the Clément operator, since $N_i(\Pi_i u)$, $i = N + 1, \dots, N_+$, is not necessarily zero for $u \in V$. Consequently, the elements adjacent to the Dirichlet boundary must be treated separately in the analysis of the interpolation error. An alternative is developed by Scott and Zhang (1990).

Example 5 (Scott–Zhang operator) The operator is introduced by Scott and Zhang (1990) similar to the Clément operator (see Example 4). In particular, the projector $\Pi_i: L^1(\omega_i) \rightarrow \mathbb{P}_{\ell-1}$ is also defined by (8). The essential difference is that ω_i (still satisfying $a^i \in \bar{\omega}_i$) is allowed to be a $(d-1)$ -dimensional face of an element K_i , and for Dirichlet boundary nodes, one chooses ω_i to be part of the Dirichlet boundary Γ_D . In this way, we obtain $N_i(\Pi_i u) = 0$ if $a^i \in \Gamma_D$. The operator preserves even nonhomogeneous Dirichlet boundary conditions $v|_{\Gamma_D} = g$ if $g \in \mathbb{P}_{\ell-1}$ and $\ell = k + 1$ in (7).

To be specific about the choice of ω_i for $a^i \notin \Gamma_D$, we recall from Scott and Zhang (1990) that $\bar{\omega}_i = K \in \mathcal{T}$ if $a^i \in \text{int } K$, and $\bar{\omega}_i \ni a^i$ is a face of some element otherwise. Note that the face is not uniquely determined if a^i does not lie in the interior of a face or an element. For an illustration and an application of this operator in a context where the nodal interpolant is not applicable, we refer to Apel, Sändig, and Whiteman (1996).

The operator can be applied to functions whose traces on $(d-1)$ -dimensional manifolds ω_i are in $L^1(\omega_i)$, that means, for $u \in W^{\ell,p}(\Omega)$ with $\ell \geq 1$ and $p = 1$, or with $\ell > 1/p$ and $p > 1$. Consequently, it requires more regularity than the Clément operator, but, in general, less than the nodal interpolant.

Finally, Verfürth (1999a) remarks that in certain interpolation error estimates that are valid for both the Scott–Zhang and the Clément operators, the constant is smaller for the Clément operator.

4 THE DENY–LIONS LEMMA

In this section, we discuss a result from functional analysis that turns out to be a classical approximation result, the Deny–Lions lemma (Deny and Lions, 1953/54), which is an essential ingredient of many error estimates in the finite element theory. It essentially states that the $W^{\ell,p}(G)$ -seminorm is a norm in the quotient space $W^{\ell,p}(G)/\mathbb{P}_{\ell-1}$. We formulate it for domains G of unit size $\text{diam } G = 1$.

Lemma 1 (Deny and Lions) Let the domain $G \subset \mathbb{R}^d$, $\text{diam } G = 1$, be star-shaped with respect to a ball $B \subset G$, and let $\ell \geq 1$ be an integer and $p \in [1, \infty]$ real. For each $u \in W^{\ell,p}(G)$, there is a $w \in \mathbb{P}_{\ell-1}$ such that

$$\|u - w\|_{W^{\ell,p}(G)} \leq C \|u\|_{W^{\ell,p}(G)} \quad (11)$$

where the constant C depends only on d , ℓ , and $p := \text{diam } G / \text{diam } B = 1 / \text{diam } B$.

One can find different versions of the lemma and its proof in the literature. Instead of giving one of them in full

detail, we sketch some of them, hereby elucidating some important points.

A classical proof is to choose a basis $\{\sigma_\alpha\}_{|\alpha| \leq \ell-1}$ of $\mathbb{P}_{\ell-1}$ and to prove that $\|u\|_{W^{\ell,p}(G)} + \sum_{|\alpha| \leq \ell-1} |\sigma_\alpha(u)|$ defines a norm in $W^{\ell,p}(G)$, which is equivalent to $\|u\|_{W^{\ell,p}(G)}$. Determining $w \in \mathbb{P}_{\ell-1}$ by $\sigma_\alpha(u - w) = 0$ for all $\alpha: |\alpha| \leq \ell-1$ leads to (11). For $\ell = 1$, there is only one functional σ to be used, typically $\sigma(u) := |G|^{-1} \int_G u$. For $\ell \geq 2$, one can take the nodal variables N_ξ of a simplicial Lagrange element $(S, \mathbb{P}_{\ell-1}, N_\xi)$ with $S \subset G$ (Braess, 1997) or $\sigma_\alpha(u) := |G|^{-1} \int_G D^\alpha u$ (Bramble and Hilbert, 1970). The proof is based on the compact embedding of $W^{1,p}(G)$ in $L^p(G)$ and has the disadvantage that it only ensures that the constant is independent of u , but it can depend on all parameters, in particular, on the shape of G . The result is useful when applied only on a reference element $G = \hat{K}$. Dobrowolski (1998) uses $\sigma_\alpha(u) := |G|^{-1} \int_G D^\alpha u$ as well and obtains with a different proof that the constant is independent of the shape of G (in particular also independent of γ) but he needs the assumption that G is convex.

Dupont and Scott (1980) choose w to be the averaged Taylor polynomial (also called Sobolev polynomial)

$$T_B^\ell u(x) := \int_B T_\gamma^\ell u(x) \phi(y) dy \in \mathbb{P}_{\ell-1}$$

where the Taylor polynomial of order $\ell-1$ evaluated at y is given by

$$T_\gamma^\ell u(x) := \sum_{|\alpha| \leq \ell-1} \frac{1}{\alpha!} D^\alpha u(y) (x - y)^\alpha \in \mathbb{P}_{\ell-1}$$

and where B is the ball from the lemma, and $\phi(\cdot)$ is a smooth cut-off function with support \bar{B} and $\int_{\mathbb{R}^d} \phi = 1$; see also Brenner and Scott (1994, Section 4.1). This polynomial has several advantages, among them the property

$$D^\alpha T_B^\ell u = T_B^{\ell-|\alpha|} D^\alpha u \quad \forall u \in W^{1,1}(B) \quad (12)$$

which will lead to simplifications later. This choice of w also allows the characterization of the constant in (11) as stated in Lemma 1. Moreover, several extensions can be made; so the domain may be generalized to a union of overlapping domains that are each star-shaped with respect to a ball, and the Sobolev index ℓ may be noninteger (see the original paper by Dupont and Scott, 1980).

Verfürth (1999b) defines in a recursive manner the projector $P_G^\ell: H^\ell(\Omega) \rightarrow \mathbb{P}_\ell$,

$$P_G^\ell(u) := \sum_{|\alpha|=\ell} \frac{x^\alpha}{\alpha!} \frac{1}{|G|} \int_G D^\alpha u,$$

$$P_G^{\ell-1}(u) := P_G^\ell(u) + \sum_{|\alpha|=\ell-1} \frac{x^\alpha}{\alpha!} \frac{1}{|G|} \int_G [D^\alpha(u - P_G^\ell(u))],$$

$$k = \ell, \ell-1, \dots, 1,$$

$$P_G^0 := P_G^0(u) \quad (13)$$

which also commutes with differentiation in the sense of (12) and allows to prove (11) for $w = P_G^{\ell-1}u$ with a constant C , depending only on ℓ and $p \in [2, \infty]$. The restriction $p \geq 2$ is outweighed by the fact that for convex Ω the constant C does not depend on the parameter $\gamma := \text{diam } G / \text{diam } B$.

5 LOCAL ERROR ESTIMATES FOR THE NODAL INTERPOLANT

5.1 Isotropic elements

Interpolation error estimates can be proved on the basis of Lemma 1. To elaborate how the estimates depend on the element size $h_K := \text{diam } K \leq C \Omega_K$, the isotropic element K is mapped to a domain of unit size. In simple cases, when $\mathbb{P}_{\ell-1} \subset \mathcal{P}_K$ (this is, in general, not satisfied for isoparametric elements), one can just scale via

$$x = h_K \bar{x}$$

where $x, \bar{x} \in \mathbb{R}^d$. This transformation maps K to a similar element \bar{K} of unit size, $\text{diam } \bar{K} = 1$; it also maps $\mathbb{P}_{\ell-1}$ to $\mathbb{P}_{\ell-1}$, $u \in W^{\ell,p}(K)$ to $\bar{u} \in W^{\ell,p}(\bar{K})$ and the nodal interpolant $I_K u$ to $I_{\bar{K}} \bar{u}$. So, we get with $I_K \bar{u} = \bar{u}$ for all $\bar{u} \in \mathbb{P}_{\ell-1}$ and \bar{u} according to Lemma 1

$$\begin{aligned} \|u - I_K u\|_{W^{m,p}(K)} &= h_K^{d/p} h_K^{-m} \|\bar{u} - I_{\bar{K}} \bar{u}\|_{W^{m,p}(\bar{K})}, \\ \|\bar{u} - I_{\bar{K}} \bar{u}\|_{W^{m,p}(\bar{K})} &= \|(\bar{u} - \bar{w}) - I_{\bar{K}}(\bar{u} - \bar{w})\|_{W^{m,p}(\bar{K})} \\ &\leq \|\bar{u} - \bar{w}\|_{W^{m,p}(\bar{K})} + \|I_{\bar{K}}(\bar{u} - \bar{w})\|_{W^{m,p}(\bar{K})} \\ &\leq (1 + \|I_{\bar{K}}\|_{W^{\ell,p}(\bar{K}) \rightarrow W^{m,p}(\bar{K})}) \\ &\quad \times \|\bar{u} - \bar{w}\|_{W^{\ell,p}(\bar{K})} \\ &\leq (1 + \|I_{\bar{K}}\|_{W^{\ell,p}(\bar{K}) \rightarrow W^{m,p}(\bar{K})}) \cdot C \|\bar{u}\|_{W^{\ell,p}(\bar{K})} \end{aligned}$$

Scaling back, we obtain for $m = 0, \dots, \ell$

$$\|u - I_K u\|_{W^{m,p}(K)} \leq C(1 + \|I_{\bar{K}}\|_{W^{\ell,p}(\bar{K}) \rightarrow W^{m,p}(\bar{K})}) \times h_K^{\ell-m} \|u\|_{W^{\ell,p}(K)}$$

The operator norm $\|\mathbb{I}_{\tilde{K}}\|_{W^{t,p}(\tilde{K}) \rightarrow W^{s,p}(\tilde{K})} := \sup_{\tilde{u} \in W^{t,p}(\tilde{K})} \|\mathbb{I}_{\tilde{K}} \tilde{u}\|_{W^{s,p}(\tilde{K})} / \|\tilde{u}\|_{W^{t,p}(\tilde{K})}$ can be bounded by using

$$\|\mathbb{I}_{\tilde{K}} \tilde{u}\|_{W^{s,p}(\tilde{K})} = \left\| \sum_{i=1}^n N_{i,\tilde{K}}(\tilde{u}) \Phi_{i,\tilde{K}} \right\|_{W^{s,p}(\tilde{K})} \leq \sum_{i=1}^n |N_{i,\tilde{K}}(\tilde{u})| \|\Phi_{i,\tilde{K}}\|_{W^{s,p}(\tilde{K})}$$

It remains to be proved that $\|\Phi_{i,\tilde{K}}\|_{W^{s,p}(\tilde{K})} \leq C$ and $|N_{i,\tilde{K}}(\tilde{u})| \leq C \|\tilde{u}\|_{W^{t,p}(\tilde{K})}$. To ensure that these constants are independent of the form of \tilde{K} , Brenner and Scott (1994) transform $(\tilde{K}, \mathcal{P}_{\tilde{K}}, \mathcal{N}_{\tilde{K}})$ to the reference element $(\hat{K}, \mathcal{P}_{\hat{K}}, \mathcal{N}_{\hat{K}})$, which was introduced in Assumption 1.

Other references suggest instead that the interpolation error estimates by transforming immediately to the reference element \hat{K} be proved. We will discuss this approach in the remaining part of this subsection. Recall from Section 2 that we assume that for each element $K \in \mathcal{T}$ there is a bijective mapping $F_K: \hat{x} \in \hat{K} \mapsto x = F_K(\hat{x}) \in K$, which maps \hat{K} to K . The following lemma provides transformation formulae for seminorms of functions if F_K is affine.

Lemma 2. Let $F_K(\hat{x}) = A\hat{x} + a$ be an affine mapping with $K = F_K(\hat{K})$. If $\hat{u} \in W^{m,q}(\hat{K})$, then $u = \hat{u} \circ F_K^{-1} \in W^{m,q}(K)$ and

$$|u|_{W^{m,q}(K)} \leq C |K|^{1/q} \Omega_K^{-m} |\hat{u}|_{W^{m,q}(\hat{K})} \quad (14)$$

If $u \in W^{t,p}(K)$, then $\hat{u} = u \circ F_K \in W^{t,p}(\hat{K})$ and

$$|\hat{u}|_{W^{t,p}(\hat{K})} \leq C |K|^{-1/p} h_K^t |u|_{W^{t,p}(K)} \quad (15)$$

The constants depend on the shape and size of \hat{K} .

Proof. We follow Ciarlet (1978). By examining the affine mapping, we get $\hat{\nabla} \hat{u} = A^T \nabla u$ and thus

$$|u|_{W^{m,q}(K)} \leq C |K|^{1/q} \|A^{-1}\|_2^m |\hat{u}|_{W^{m,q}(\hat{K})} \\ |\hat{u}|_{W^{t,p}(\hat{K})} \leq C |K|^{-1/p} \|A\|_2^t |u|_{W^{t,p}(K)}$$

The factor with the power of $|K|$ comes from the Jacobi determinant of the transformation. This determinant is equal to the ratio of the areas of K and \hat{K} . The norm of A can be estimated by considering the transformation of the largest sphere \hat{S} contained in \hat{K} . For all $\hat{x} \in \hat{K}$ with $|\hat{x}| = \Omega_K = \text{diam } \hat{S}$, there are two points $\hat{y}, \hat{z} \in \hat{S}$ such that $\hat{x} = \hat{y} - \hat{z}$. By observing that $|A\hat{x}| = |(A\hat{y} + a) - (A\hat{z} + a)| = |\hat{y} - \hat{z}| \leq h_K$, we get

$$\|A\|_2 := \sup_{|\hat{x}|=\Omega_K} \frac{|A\hat{x}|}{|\hat{x}|} \leq \frac{h_K}{\Omega_K}$$

and analogously $\|A^{-1}\|_2 \leq h_K/\Omega_K$. This finishes the proof. \square

Theorem 1. Let $(\tilde{K}, \mathcal{P}_{\tilde{K}}, \mathcal{N}_{\tilde{K}})$ be a reference element with

$$\mathcal{P}_{\ell-1} \subset \mathcal{P}_{\tilde{K}} \quad (16)$$

$$\mathcal{N}_{\tilde{K}} \subset (C^t(\tilde{K}))' \quad (17)$$

Assume that $(K, \mathcal{P}_K, \mathcal{N}_K)$ is affine equivalent to $(\tilde{K}, \mathcal{P}_{\tilde{K}}, \mathcal{N}_{\tilde{K}})$. Let $u \in W^{\ell,p}(K)$ with $\ell \in \mathbb{N}$, $p \in [1, \infty]$, such that

$$W^{\ell,p}(K) \hookrightarrow C^t(K), \quad \text{i.e. } \ell > s + \frac{d}{p} \quad (18)$$

and let $m \in \{0, \dots, \ell-1\}$ and $q \in [1, \infty]$ be such that

$$W^{\ell,p}(K) \hookrightarrow W^{m,q}(\hat{K}) \quad (19)$$

Then the estimate

$$|u - \mathbb{I}_K u|_{W^{m,q}(K)} \leq C |K|^{1/q-1/p} h_K^m \Omega_K^{-m} |u|_{W^{\ell,p}(K)} \quad (20)$$

holds.

Proof. From (17) and (18), we obtain $|N_{i,\tilde{K}}(\hat{u})| \leq C \|\hat{u}\|_{C^t(\tilde{K})} \leq C \|\hat{u}\|_{W^{t,p}(\tilde{K})}$ and, thus, with $\|\Phi_{i,\tilde{K}}\|_{W^{m,q}(\tilde{K})} \leq C$, the boundedness of the interpolation operator,

$$|\mathbb{I}_{\tilde{K}} \hat{u}|_{W^{m,q}(\tilde{K})} = \left| \sum_{i=1}^n N_{i,\tilde{K}}(\hat{u}) \Phi_{i,\tilde{K}} \right|_{W^{m,q}(\tilde{K})} \leq \sum_{i=1}^n |N_{i,\tilde{K}}(\hat{u})| \|\Phi_{i,\tilde{K}}\|_{W^{m,q}(\tilde{K})} \leq C \|\hat{u}\|_{W^{t,p}(\tilde{K})}$$

where the constant depends not only on \hat{K} , s , m , q , ℓ , and p but also on $\mathcal{N}_{\tilde{K}}$. The embedding (19) yields

$$|\hat{u}|_{W^{m,q}(\tilde{K})} \leq C \|\hat{u}\|_{W^{t,p}(\tilde{K})}$$

Combining these estimates, choosing $\hat{u} \in \mathbb{P}_{\ell-1}$ according to Lemma 1, and using $\hat{u} = \mathbb{I}_{\tilde{K}} \hat{u}$ due to (16), we get

$$|\hat{u} - \mathbb{I}_{\tilde{K}} \hat{u}|_{W^{m,q}(\tilde{K})} = |(\hat{u} - \hat{u}) - \mathbb{I}_{\tilde{K}}(\hat{u} - \hat{u})|_{W^{m,q}(\tilde{K})} \leq C \|\hat{u} - \hat{u}\|_{W^{t,p}(\tilde{K})} \leq C |\hat{u}|_{W^{t,p}(\tilde{K})} \quad (21)$$

By transforming this estimate to K (using Lemma 2), we obtain the desired result. \square

Note that this theorem restricts for simplicity to affine elements, but is valid not only for Lagrangian elements but also for other types, including Hermite elements.

Corollary 1. For isotropic elements, we obtain, in particular,

$$|u - \mathbb{I}_K u|_{W^{m,q}(K)} \leq C |K|^{1/q-1/p} h_K^{\ell-m} |u|_{W^{\ell,p}(K)} \quad (22)$$

Remark 6. Interpolation error estimates can also be proved for functions from weighted Sobolev spaces, for example,

$$H^{2,\alpha}(G) := \{u \in W^{1,2}(G) : r^\alpha D^\beta u \in L^2(G) \forall \beta : |\beta| = 2\}$$

where r is the distance to some point $x \in \overline{G} \subset \mathbb{R}^2$, and

$$|u|_{H^{2,\alpha}(G)}^2 := \sum_{|\beta|=2} \|r^\alpha D^\beta u\|_{L^2(G)}^2$$

$$\|u\|_{H^{2,\alpha}(G)}^2 := \|u\|_{W^{1,2}(G)}^2 + |u|_{H^{2,\alpha}(G)}^2$$

Grisvard (1985) shows in Lemma 8.4.1.3 the analog to the Deny–Lions lemma. For each $u \in H^{2,\alpha}(G)$ with $\alpha < 1$, there is a $w \in \mathbb{P}_1$ such that

$$\|u - w\|_{H^{2,\alpha}(G)} \leq C(G) |u|_{H^{2,\alpha}(G)}$$

The interpolation error estimate

$$|u - \mathbb{I}_K u|_{W^{1,2}(K)} \leq C h_K^2 \Omega_K^{-1-\alpha} |u|_{H^{2,\alpha}(K)}$$

is then proved in Lemma 8.4.1.4 for triangles K . This result can be applied in the proof of mesh-grading techniques for singular solutions, where the singularity comes from corners in the domain $\Omega \subset \mathbb{R}^2$.

Second derivatives of an affine transformation F_K vanish. This leads to the special structure of the relations (14) and (15), where no low-order derivatives of \hat{u} and u , respectively, appear on the right-hand sides. This is no longer valid for nonaffine transformations. In the case that

$$|\hat{F}_K^{\alpha}| \leq C h_K^{|\alpha|} \quad \forall \alpha : |\alpha| \leq \ell \quad (23)$$

we obtain

$$|\hat{u}|_{W^{\ell,p}(\hat{K})} \leq C |K|^{-1/p} h_K^\ell |u|_{W^{\ell,p}(K)} \quad (24)$$

which is weaker than (15), but is still sufficient for our purposes. The assumption (23) is satisfied when F_K differs only slightly from an affine mapping.

However, Estimate (23) is not valid for general quadrilateral meshes. Therefore, the theory has to be refined. For $\mathcal{P}_K = \mathcal{Q}_K$, this case can be treated with a sharper version of the Deny–Lions lemma: for each $u \in W^{\ell,p}(G)$ there is

a $w \in \mathcal{Q}_{\ell-1}$ such that

$$\|u - w\|_{W^{\ell,p}(G)} \leq C \left(\sum_{i=1}^d \left\| \frac{\partial^t u}{\partial x_i^t} \right\|_{L^p(G)} \right)^{1/p}$$

(see Bramble and Hilbert, 1971). For shape-regular elements (in the sense of Ciarlet/Raviart or Girault/Raviart, see Example 3) one can then prove (22).

Remark 7. Some results are weaker for general shape-regular quadrilateral elements than for (at least asymptotically) affine elements. For example, Arnold, Boffi, and Falk (2002) have shown for quadrilateral serendipity elements (here $\mathcal{P}_K = \mathcal{Q}_K' := \mathbb{P}_K \oplus \text{span}\{\hat{x}_1^2 \hat{x}_2, \hat{x}_1 \hat{x}_2^2\}$) that

$$|u - \mathbb{I}_K u|_{W^{m,2}(K)} \leq C h_K^{(k/2)+1-m} |u|_{W^{k+1,2}(K)}, \quad m = 0, 1$$

is sharp for general quadrilateral meshes, whereas for asymptotically parallelogram meshes, we get

$$|u - \mathbb{I}_K u|_{W^{m,2}(K)} \leq C h_K^{k+1-m} |u|_{W^{k+1,2}(K)}, \quad m = 0, 1$$

5.2 Anisotropic elements

Anisotropic elements are characterized by a large aspect ratio $\gamma_K := h_K/\Omega_K$. Estimate (20) can also be reformulated as

$$|u - \mathbb{I}_K u|_{W^{m,q}(K)} \leq C |K|^{1/q-1/p} h_K^{\ell-m} \gamma_K^m |u|_{W^{\ell,p}(K)}$$

which means that the quality of this estimate deteriorates if $m \geq 1$ and $\gamma_K \gg 1$. Let us examine whether the reason for this deterioration is formed by the anisotropic element (indicating that anisotropic elements should be avoided) or by the sharpness of the estimates (indicating that the estimates should be improved).

Example 6. Consider the triangle K with the nodes $(-h, 0)$, $(h, 0)$, and $(0, \varepsilon h)$ and interpolate the function $u(x_1, x_2) = x_1^2$ in the vertices with polynomials of degree one. Then $\mathbb{I}_K u = h^2 - \varepsilon^{-1} h x_2$ and

$$\frac{|u - \mathbb{I}_K u|_{W^{1,2}(K)}}{|u|_{W^{2,2}(K)}} = \left(\frac{2h^4 \left(\frac{1}{3}\varepsilon + \frac{1}{6}\varepsilon^{-1} \right)}{4h^2\varepsilon} \right)^{1/2} = h \left(\frac{1}{3} + \frac{1}{6}\varepsilon^{-2} \right)^{1/2} = c_\varepsilon h$$

with $c_\varepsilon \rightarrow \infty$ for $\varepsilon \rightarrow 0$ and $c_\varepsilon \geq C\gamma_K$. We find that Estimate (20) can, in general, not be improved essentially and that (22) is not valid. Estimate (20) can be improved only slightly by investigating in more detail the transformation from Lemma 2 (see e.g. Formaggia and Perotto, 2001).

Example 7. Consider now the triangle with the nodes $(0, 0)$, $(h, 0)$, and $(0, \varepsilon h)$ and interpolate again the function $u(x_1, x_2) = x_1^2$ in \mathbb{P}_1 . We get $I_K u = hx_1$ and

$$\frac{|u - I_K u|_{W^{1,2}(K)}}{|u|_{W^{2,2}(K)}} = \left(\frac{\frac{1}{2} \varepsilon h^4}{2 \varepsilon h^2} \right)^{1/2} = \frac{1}{\sqrt{12}} h$$

where the constant is independent of ε . Estimate (22) is valid, although the element is anisotropic for small ε .

From the two examples, we can learn that the aspect ratio is not the right quantity to characterize elements that yield an interpolation error estimate of quality (22). Syngé (1957) proved for triangles K and $\mathcal{P}_K = \mathbb{P}_1$ that

$$|u - I_K u|_{W^{1,\infty}(K)} \leq Ch_K |u|_{W^{2,\infty}(K)}$$

with a constant that depends linearly on $(\cos(1/2)\alpha)^{-1}$, where α is the maximal angle in K . The maximal angle condition was found (see also the comment in Example 2).

We will now elaborate that Estimate (22) cannot be obtained from (21) by just treating the transformation F_K more carefully (compare also Apel, 1999a, Example 2.1). To do so, we consider again the triangle K from Example 7 with $\mathcal{P}_K = \mathbb{P}_1$. The transformation to the reference element \hat{K} with nodes $(0, 0)$, $(1, 0)$, and $(0, 1)$ is done via $x_1 = h\hat{x}_1$, $x_2 = \varepsilon h\hat{x}_2$. Transforming (21) in the special case $p = q = 2$, $m = 1$, $\ell = 2$ leads to

$$\begin{aligned} & h \left(\left\| \frac{\partial(u - I_K u)}{\partial x_1} \right\|_{L^2(K)}^2 + \varepsilon^2 \left\| \frac{\partial(u - I_K u)}{\partial x_2} \right\|_{L^2(K)}^2 \right)^{1/2} \\ & \leq Ch^2 \left(\left\| \frac{\partial^2 u}{\partial x_1^2} \right\|_{L^2(K)}^2 + \varepsilon^2 \left\| \frac{\partial^2 u}{\partial x_1 \partial x_2} \right\|_{L^2(K)}^2 + \varepsilon^4 \left\| \frac{\partial^2 u}{\partial x_2^2} \right\|_{L^2(K)}^2 \right)^{1/2} \end{aligned}$$

which can be simplified to

$$\begin{aligned} \left\| \frac{\partial(u - I_K u)}{\partial x_1} \right\|_{L^2(K)} & \leq Ch |u|_{W^{2,2}(K)} \\ \left\| \frac{\partial(u - I_K u)}{\partial x_2} \right\|_{L^2(K)} & \leq C \varepsilon^{-1} h |u|_{W^{2,2}(K)} \end{aligned} \quad (25)$$

but the independence of ε^{-1} is not obtainable in (25). This factor could only be avoided if we proved on the reference element the sharper estimate

$$\left\| \frac{\partial(\hat{u} - I_{\hat{K}} \hat{u})}{\partial \hat{x}_2} \right\|_{L^2(\hat{K})} \leq C \left(\left\| \frac{\partial^2 \hat{u}}{\partial \hat{x}_1 \partial \hat{x}_2} \right\|_{L^2(\hat{K})}^2 + \left\| \frac{\partial^2 \hat{u}}{\partial \hat{x}_2^2} \right\|_{L^2(\hat{K})}^2 \right)^{1/2}$$

The following lemma from Apel and Dobrowolski (1992) (see also Apel, 1999a, Lemma 2.2) reduces the proof to finding functionals σ_i with certain properties.

Lemma 3. Let $I_K: C^1(\hat{K}) \rightarrow \mathcal{P}_K$ be a linear operator and assume that $\mathbb{P}_1 \subset \mathcal{P}_K$. Fix $m, \ell \in \mathbb{N}_0$ and $p, q \in [1, \infty]$ such that $0 \leq m \leq \ell \leq k+1$ and

$$W^{\ell-m}(\hat{K}) \hookrightarrow L^q(\hat{K}) \quad (26)$$

Consider a multi-index γ with $|\gamma| = m$ and define $J := \dim \hat{D}^\gamma \mathcal{P}_K$. Assume that there are linear functionals σ_i , $i = 1, \dots, J$, such that

$$\sigma_i \in (W^{\ell-m,p}(\hat{K}))', \quad \forall i = 1, \dots, J \quad (27)$$

$$\sigma_i(\hat{D}^\gamma(\hat{u} - I_{\hat{K}} \hat{u})) = 0 \quad \forall i = 1, \dots, J,$$

$$\forall \hat{u} \in C^1(\hat{K}): \hat{D}^\gamma \hat{u} \in W^{\ell-m,p}(\hat{K}) \quad (28)$$

$$\hat{u} \in \mathcal{P}_K \text{ and } \sigma_i(\hat{D}^\gamma \hat{u}) = 0 \quad \forall i = 1, \dots, J,$$

$$\Rightarrow \hat{D}^\gamma \hat{u} = 0 \quad (29)$$

Then the error can be estimated for all $\hat{u} \in C^1(\hat{K})$ with $\hat{D}^\gamma \hat{u} \in W^{\ell-m,p}(\hat{K})$ by

$$\|\hat{D}^\gamma(\hat{u} - I_{\hat{K}} \hat{u})\|_{L^q(\hat{K})} \leq C |\hat{D}^\gamma \hat{u}|_{W^{\ell-m,p}(\hat{K})} \quad (30)$$

Proof. The proof is based on two ingredients. First, we conclude from (12) and the Deny–Lions lemma 1 that $\hat{u} = T_B^* \hat{u} \in \mathbb{P}_{\ell-1} \subset \mathcal{P}_K$ satisfies

$$\hat{u}_\gamma := \hat{D}^\gamma \hat{u} = T_B^{-|\gamma|} \hat{D}^\gamma \hat{u} \in \mathbb{P}_{\ell-m}$$

and

$$\|\hat{D}^\gamma \hat{u} - \hat{u}_\gamma\|_{W^{\ell-m,p}(\hat{K})} \leq C |\hat{D}^\gamma \hat{u}|_{W^{\ell-m,p}(\hat{K})} \quad (31)$$

Second, we see that $\hat{D}^\gamma(\hat{u} - I_{\hat{K}} \hat{u}) \in \hat{D}^\gamma \mathcal{P}_K$. Moreover, $\sum_{i=1}^J |\sigma_i(\cdot)|$ and $\|\cdot\|_{L^q(\hat{K})}$ are equivalent norms in $\hat{D}^\gamma \mathcal{P}_K$. Therefore, we get with (28) and (27)

$$\begin{aligned} \|\hat{D}^\gamma(\hat{u} - I_{\hat{K}} \hat{u})\|_{L^q(\hat{K})} & \leq C \sum_{i=1}^J |\sigma_i(\hat{D}^\gamma(\hat{u} - I_{\hat{K}} \hat{u}))| \\ & = C \sum_{i=1}^J |\sigma_i(\hat{D}^\gamma(\hat{u} - \hat{u}))| \\ & \leq C \|\hat{D}^\gamma(\hat{u} - \hat{u}_\gamma)\|_{W^{\ell-m,p}(\hat{K})} \end{aligned}$$

Consequently, we obtain with (26) and (31)

$$\begin{aligned} \|\hat{D}^\gamma(\hat{u} - I_{\hat{K}} \hat{u})\|_{L^q(\hat{K})} & \leq \|\hat{D}^\gamma(\hat{u} - \hat{u}_\gamma)\|_{L^q(\hat{K})} \\ & \quad + \|\hat{D}^\gamma(\hat{u}_\gamma - I_{\hat{K}} \hat{u}_\gamma)\|_{L^q(\hat{K})} \\ & \leq C \|\hat{D}^\gamma(\hat{u} - \hat{u}_\gamma)\|_{W^{\ell-m,p}(\hat{K})} \\ & \leq |\hat{D}^\gamma \hat{u}|_{W^{\ell-m,p}(\hat{K})} \end{aligned}$$

The creative part is now to find the functionals $\{\sigma_i\}_{i=1}^J$ with the properties (27)–(29).

Example 8. For Lagrangian elements and $m = 1$, the functionals are integrals over some lines (see Figure 3). One can easily check (28) and (29). The critical condition is (27), which is satisfied for $\ell = 2$ only if $d = 2$ or $p > 2$. One can indeed give an example that shows that

$$\|\hat{D}^\gamma(\hat{u} - I_{\hat{K}} \hat{u})\|_{L^2(\hat{K})} \leq C |\hat{D}^\gamma \hat{u}|_{W^{1,p}(\hat{K})}$$

does not hold for $d = 3$, $p \leq 2$ (see Apel and Dobrowolski, 1992).

Let us transform Estimate (30) to the element $K = F_K(\hat{K})$. We see easily that if $F_K(\hat{x}) = A\hat{x} + a$ with $A = \text{diag}(h_{1,K}, \dots, h_{d,K})$, we get

$$\begin{aligned} & |K|^{-1/q} h_{1,K}^{q_1} \dots h_{d,K}^{q_d} \|D^\gamma(u - I_K u)\|_{L^q(K)} \\ & \leq C |K|^{-1/p} h_{1,K}^{q_1} \dots h_{d,K}^{q_d} \|D^{\alpha+\gamma} u\|_{L^p(K)} \\ & \quad \times \sum_{|\alpha|=\ell-m} h_{1,K}^{q_1} \dots h_{d,K}^{q_d} \|D^{\alpha+\gamma} u\|_{L^p(K)} \end{aligned}$$

Dividing by $|K|^{-1/q} h_{1,K}^{q_1} \dots h_{d,K}^{q_d}$ and summing up over all γ with $|\gamma| = m$, we obtain

$$\begin{aligned} |u - I_K u|_{W^{m,p}(K)} & \leq C |K|^{1/q-1/p} \\ & \quad \times \sum_{|\alpha|=\ell-m} h_{1,K}^{q_1} \dots h_{d,K}^{q_d} |D^{\alpha} u|_{W^{m,p}(K)} \end{aligned} \quad (32)$$

Remark 8. A more detailed calculation shows that this result can also be obtained when the off-diagonal entries of $A = [a_{i,j}]_{i,j=1}^d$ are not zero but small,

$$a_{i,j} \leq C \min\{h_{i,K}, h_{j,K}\}, \quad i, j = 1, \dots, d, \quad i \neq j \quad (33)$$



Figure 3. Functionals for Lagrangian elements.

see Apel and Lube (1998). A geometrical description of two- and three-dimensional affine elements that satisfy (33) is given in terms of a *maximal angle condition* and a *coordinate system condition* in Apel (1999a).

The situation is more difficult for nonaffine elements.

Example 9. Consider the quadrilateral K with nodes $(0, 0)$, $(h_1, 0)$, (h_1, h_2) , (ε, h_2) , $0 < \varepsilon \leq h_2 \leq h_1$, which is an ε -perturbation of the (affine) rectangle $(0, h_1) \times (0, h_2)$. We have

$$F_K(\hat{x}) = \begin{pmatrix} h_1 \hat{x}_1 + \varepsilon(1 - \hat{x}_1) \hat{x}_2 \\ h_2 \hat{x}_2 \end{pmatrix}$$

and as in the affine theory

$$|u - I_K u|_{W^{1,2}(K)} \leq |K|^{1/2} \sum_{i=1}^2 h_i^{-1} \left\| \frac{\partial(\hat{u} - I_{\hat{K}} \hat{u})}{\partial \hat{x}_i} \right\|_{L^2(\hat{K})}$$

$$\leq |K|^{1/2} \sum_{i=1}^2 h_i^{-1} \left\| \frac{\partial \hat{u}}{\partial \hat{x}_i} \right\|_{L^2(\hat{K})}$$

$$\left\| \frac{\partial^2 \hat{u}}{\partial \hat{x}_1^2} \right\|_{L^2(\hat{K})} \leq C |K|^{-1/2} h_1^2 \left\| \frac{\partial^2 u}{\partial x_1^2} \right\|_{L^2(K)}$$

$$\left\| \frac{\partial^2 \hat{u}}{\partial \hat{x}_2^2} \right\|_{L^2(\hat{K})} \leq C |K|^{-1/2} h_2^2 |u|_{W^{2,2}(K)}$$

but owing to $\partial^2 x_1 / \partial \hat{x}_1 \partial \hat{x}_2 = -\varepsilon \neq 0$, we get only

$$\left\| \frac{\partial^2 \hat{u}}{\partial \hat{x}_1 \partial \hat{x}_2} \right\|_{L^2(\hat{K})} \leq C |K|^{-1/2} \left(h_1 h_2 \left\| \frac{\partial u}{\partial x_1} \right\|_{W^{1,2}(K)} + \varepsilon \left\| \frac{\partial u}{\partial x_1} \right\|_{L^2(K)} \right)$$

and, thus, by using again $\varepsilon \leq h_2 \leq h_1$

$$\begin{aligned} |u - I_K u|_{W^{1,2}(K)} & \leq C \left(\sum_{i=1}^2 h_i \left\| \frac{\partial u}{\partial x_i} \right\|_{W^{1,2}(K)} + \frac{\varepsilon}{h_2} \left\| \frac{\partial u}{\partial x_1} \right\|_{L^2(K)} \right) \end{aligned}$$

In Apel (1998), we concluded that we should allow only perturbations with $\varepsilon \leq Ch_1 h_2$, but later we found in Apel (1999a) a sharper estimate without the latter term: observing that $\mathbb{P}_1 \in \mathcal{P}_K$, we get for $w \in \mathbb{P}_1$

$$\begin{aligned} |u - I_K u|_{W^{1,2}(K)} & = |(u - w) - I_K(u - w)|_{W^{1,2}(K)} \\ & \leq C \left(\sum_{i=1}^2 h_i \left\| \frac{\partial u}{\partial x_i} \right\|_{W^{1,2}(K)} + \frac{\varepsilon}{h_2} \left\| \frac{\partial(u - w)}{\partial x_1} \right\|_{L^2(K)} \right) \end{aligned}$$

By another Deny–Lions argument, comparing (31), we get for appropriate w

$$\left\| \frac{\partial(u-w)}{\partial x_1} \right\|_{L^2(K)} \leq C \sum_{i=1}^2 h_i \left\| \frac{\partial^2 u}{\partial x_1 \partial x_i} \right\|_{L^2(K)}$$

such that

$$\|u - I_K u\|_{W^{1,2}(K)} \leq C \sum_{i=1}^2 h_i \left\| \frac{\partial u}{\partial x_i} \right\|_{W^{1,2}(K)}$$

can be proved for $\epsilon \leq Ch_2$.

The approach from the example, where a second Deny–Lions argument is used, holds also for more general quadrilateral elements K with straight edges (subparametric elements) and $P_K = Q_K$ (see Apel, 1999a).

Summarizing this subsection, we can say that the anisotropic interpolation error estimate (32) can be proved for a large class of affine and subparametric elements (for details we refer to Apel, 1999a). Also, estimates for functions from weighted Sobolev spaces have been proved; see Apel and Nicaise (1996, 1998) and Apel, Nicaise, and Schöberl (2001). The anisotropic interpolation error estimates are suited to compensate large partial derivatives $D^2 u$ by an appropriate choice of the element sizes $h_{1,K}, \dots, h_{d,K}$ in order to equilibrate the terms in the sum at the right-hand side of (32). The results can be applied to problems with anisotropic solutions; in particular, flow problems where first results on the resolution of all kinds of layers or shock fronts can be found, for example, in Peraire, Vahdati, Morgan, and Zienkiewicz (1987), Kornhuber and Rottzsch (1990), Zhou and Rannacher (1993), Zienkiewicz and Wu (1994), and Roos, Stynes, and Tobiska (1996).

6 LOCAL ERROR ESTIMATES FOR QUASI-INTERPOLANTS

Recall from Section 3.2 that we restrict ourselves here to Lagrangian elements, that is, $N_{i,K}(u) = u(a^i)$ where a^i , $i = 1, \dots, n$, are the nodes of K . The quasi-interpolants can be defined locally by

$$Q_K u := \sum_{i=1}^n N_{i,K}(\Pi_{i,K} u) \phi_{i,K}$$

with projectors $\Pi_{i,K}: L^1(\omega_{i,K}) \rightarrow \mathbb{P}_{\ell-1}$ and sets $\omega_{i,K}$ that are defined differently by Clément and Scott/Zhang (see Examples 4 and 5, respectively). The local number of degrees of freedom that defines the operator correctly is denoted by \tilde{n} . For the Scott–Zhang operator, we can use

$\tilde{n} = n$. For the Clément operator, we also have $\tilde{n} = n$ if $K \cap \Gamma_D = \emptyset$, but $\tilde{n} < n$ if K touches the Dirichlet boundary. Let $\omega_K \subset \Omega$ be the interior of a union of finite elements with $K \subset \bar{\omega}_K$ and $\omega_{i,K} \subset \omega_K$, $i = 1, \dots, \tilde{n}$; typically, one defines

$$\bar{\omega}_K := \bigcup_{K_i \in \mathcal{T}_K, K_i \cap K \neq \emptyset} K_i$$

We will prove error estimates in a uniform way for both operators and for triangles, tetrahedra, quadrilaterals, and hexahedra, but we restrict ourselves to affine isotropic elements.

To bound the interpolation error $u - Q_K u$, we need several ingredients. The first one is the inclusion $\mathbb{P}_{\ell-1} \subset \mathcal{P}_K$ which is satisfied for the affine elements mentioned above if $P_K = \mathbb{P}_\ell$ or $P_K = Q_K$ and $\ell \leq k+1$. Then we obtain

$$Q_K w = w \quad \forall w \in \mathbb{P}_{\ell-1} \quad (34)$$

because of $N_{i,K}(\Pi_{i,K} w) = N_{i,K}(w)$ for $w \in \mathbb{P}_{\ell-1}$.

Since $\Pi_{i,K} v$ is from a finite-dimensional space, we have

$$\|\Pi_{i,K} v\|_{L^\infty(\omega_{i,K})} \leq C |\omega_{i,K}|^{-1/2} \|\Pi_{i,K} v\|_{L^2(\omega_{i,K})}$$

Moreover, we get from the definition (8) with $\phi = \Pi_{i,K} u$ that

$$\begin{aligned} \|\Pi_{i,K} v\|_{L^2(\omega_{i,K})}^2 &= \int_{\omega_{i,K}} v \Pi_{i,K} v \\ &\leq \|v\|_{L^1(\omega_{i,K})} \|\Pi_{i,K} v\|_{L^\infty(\omega_{i,K})} \end{aligned}$$

that means

$$|N_{i,K}(\Pi_{i,K} v)| \leq \|\Pi_{i,K} v\|_{L^\infty(\omega_{i,K})} \leq C |\omega_{i,K}|^{-1} \|v\|_{L^1(\omega_{i,K})} \quad (35)$$

If $\omega_{i,K}$ is d -dimensional, we conclude

$$|N_{i,K}(\Pi_{i,K} v)| \leq C |K|^{-1} \|v\|_{L^1(K)}$$

If $\omega_{i,K}$ is $(d-1)$ -dimensional, namely, a face of an element $K_i \subset \omega_K$, we need to apply the trace theorem on the reference element. By transforming to K_i , we obtain

$$\begin{aligned} \|v\|_{L^1(\omega_{i,K})} &\leq C |\omega_{i,K}| |K_i|^{-1} (\|v\|_{L^1(K_i)} + h_{K_i} \|v\|_{W^{1,1}(K_i)}) \\ |N_{i,K}(\Pi_{i,K} v)| &\leq C |K|^{-1} (\|v\|_{L^1(\omega_K)} + h_K \|v\|_{W^{1,1}(\omega_K)}) \end{aligned}$$

where we used the fact that adjacent isotropic elements are of comparable size: if $K_i \cap K_j \neq \emptyset$ then $h_{K_i} \leq C h_{K_j}$ with a constant C_T of moderate size, for example, $C_T = 2$. We

are now able to bound the norm of Q_K by

$$\begin{aligned} |Q_K v|_{W^{m,q}(K)} &= \left| \sum_{i=1}^n N_{i,K}(\Pi_{i,K} v) \phi_{i,K} \right|_{W^{m,q}(K)} \\ &\leq \sum_{i=1}^n |N_{i,K}(\Pi_{i,K} v)| |\phi_{i,K}|_{W^{m,q}(K)} \\ &\leq C |K|^{-1} \sum_{i=1}^n h_K^{1-m} \|v\|_{W^{1,1}(\omega_K)} |K|^{1/q} h_K^{-m} \\ &\leq C |K|^{1/q-1/p} \sum_{j=0}^{\ell} h_K^{j-m} \|v\|_{W^{j,p}(\omega_K)} \quad (36) \end{aligned}$$

with $\ell \geq 0$ for the Clément operator and $\ell \geq 1$ for the Scott–Zhang operator. We have also used the Hölder inequality $\|v\|_{L^1(K)} \leq \|1\|_{L^{p'}(K)} \|v\|_{L^p(K)} = |K|^{1-1/p} \|v\|_{L^p(K)}$.

By transforming the embedding theorem $W^{\ell,p}(\hat{K}) \hookrightarrow W^{m,q}(\hat{K})$, which we assume to hold, we get

$$\|v\|_{W^{m,q}(K)} \leq C |K|^{1/q-1/p} \sum_{j=0}^{\ell} h_K^{j-m} \|v\|_{W^{j,p}(K)} \quad (37)$$

where we used the formulae from Lemma 2 and $h_K \leq C \bar{\omega}_K$.

The next ingredient we need is a version of the Deny–Lions lemma 1. By the scaling $x = h_K \bar{x} \in \mathbb{R}^d$, we transform ω_K to $\bar{\omega}_K$, which is an isotropic domain with diameter of order 1. Thus, for any $\bar{u} \in W^{\ell,p}(\bar{\omega}_K)$, there is a polynomial $\bar{w} \in \mathbb{P}_{\ell-1}$ such that

$$\|\bar{u} - \bar{w}\|_{W^{\ell,p}(\bar{\omega}_K)} \leq C \|\bar{u}\|_{W^{\ell,p}(\bar{\omega}_K)}$$

Scaling back, we obtain

$$\sum_{j=0}^{\ell} h_K^j \|u - w\|_{W^{j,p}(\omega_K)} \leq C h_K^\ell \|u\|_{W^{\ell,p}(\omega_K)} \quad (38)$$

Finally, for Dirichlet nodes a^i , we need a sharper estimate for $|N_{i,K}(\Pi_{i,K} u)|$ than (35). Consider an element K with a Dirichlet node a^i . Then there is a face $F_i \subset \Gamma_D$ at the Dirichlet part of the boundary with $a^i \in F_i$, and an element $K_i \subset \omega_K$ with $F_i \subset \partial K_i$. By using the inverse inequality, the identity $u|_{F_i} = 0$, and the trace theorem, we get for $\ell \geq 1$

$$\begin{aligned} |N_{i,K}(\Pi_{i,K} u)| &= |\Pi_{i,K} u(a^i)| \leq \|\Pi_{i,K} u\|_{L^\infty(F_i)} \\ &\leq C |F_i|^{-1} \|\Pi_{i,K} u\|_{L^1(F_i)} \\ &= C |F_i|^{-1} \|u - \Pi_{i,K} u\|_{L^1(F_i)} \end{aligned}$$

$$\leq C |K_i|^{-1/p} \sum_{j=0}^{\ell} h_{K_i}^j \|u - \Pi_{i,K} u\|_{W^{j,p}(K_i)}$$

Since (1) $K_i \subset \omega_{i,K}$ and $\omega_{i,K}$ is isotropic, (2) $\Pi_{i,K}$ is bounded in $W^{1,p}(\omega_{i,K})$, and (3) $\Pi_{i,K}$ preserves polynomials $w \in \mathbb{P}_{\ell-1}$, we obtain in analogy to (38)

$$\begin{aligned} \sum_{j=0}^{\ell} h_{K_i}^j \|u - \Pi_{i,K} u\|_{W^{j,p}(K_i)} &= \sum_{j=0}^{\ell} h_{K_i}^j \|(u-w) - \Pi_{i,K}(u-w)\|_{W^{j,p}(K_i)} \\ &\leq C \sum_{j=0}^{\ell} h_{K_i}^j \|u-w\|_{W^{j,p}(\omega_K)} \\ &\leq C h_K^\ell \|u\|_{W^{\ell,p}(\omega_K)} \end{aligned}$$

thus, we have

$$|N_{i,K}(\Pi_{i,K} u)| \leq C |K|^{-1/p} h_K^\ell \|u\|_{W^{\ell,p}(\omega_K)} \quad (39)$$

Note that this estimate holds also for $\ell = 0$ because of (35).

With these prerequisites, we obtain the final result.

Theorem 2. Let \mathcal{T} be an isotropic triangulation. Assume that each element $(F, \mathcal{P}_F, \mathcal{N}_F)$ is affine equivalent to $(\hat{K}, \mathcal{P}_{\hat{K}}, \mathcal{N}_{\hat{K}})$ with $\mathbb{P}_{\ell-1} \subset \mathcal{P}_{\hat{K}}$ and $\mathcal{N}_{\hat{K}} = \{N_{i,\hat{K}}\}_{i=1}^{\tilde{n}}$, where $N_{i,\hat{K}}(u) = u(a^i)$ and a^i are nodes. Let $u \in W^{\ell,p}(\omega_K)$, ω_K from (10), $\ell \geq 0$ for the Clément interpolant and $\ell \geq 1$ for the Scott–Zhang interpolant, $p \in [1, \infty]$. The numbers $m \in \{0, \dots, \ell-1\}$ and $q \in [1, \infty]$ are chosen such that $W^{\ell,p}(\hat{K}) \hookrightarrow W^{m,q}(\hat{K})$. Then the estimate

$$\|u - Q_K u\|_{W^{m,q}(K)} \leq C |K|^{1/q-1/p} h_K^{\ell-m} \|u\|_{W^{\ell,p}(\omega_K)}$$

holds.

Proof. Consider first the case that Q_T is the Scott–Zhang operator or, if Q_T is the Clément operator, that K does not touch the Dirichlet part of the boundary. With (34), (36), (37), we obtain for $w \in \mathbb{P}_{\ell-1}$ from (38)

$$\begin{aligned} \|u - Q_K u\|_{W^{m,q}(K)} &= \|(u-w) - Q_K(u-w)\|_{W^{m,q}(K)} \\ &\leq \|u-w\|_{W^{m,q}(K)} + |Q_K(u-w)|_{W^{m,q}(K)} \\ &\leq C |K|^{1/q-1/p} h_K^{\ell-m} \sum_{j=0}^{\ell} h_K^j \|u-w\|_{W^{j,p}(\omega_K)} \end{aligned}$$

By using (38), we obtain the desired result.

In the remaining case, we consider the Clément operator and an element K with nodes a^i , $i = 1, \dots, n$, where the nodes with $i = \tilde{n} + 1, \dots, n$ are at the boundary. Then we

write

$$|u - Q_K u|_{W^{m,p}(K)} \leq \left| u - \sum_{i=1}^n N_{i,K}(\Pi_{i,K} u) \phi_{i,K} \right|_{W^{m,p}(K)} + \sum_{i=0}^n |N_{i,K}(\Pi_{i,K} u)| |\phi_{i,K}|_{W^{m,p}(K)}$$

The first term at the right-hand side has just been estimated; the remaining terms are bounded by (39) and $|\phi_{i,K}|_{W^{m,p}(K)} \leq C|K|^{1/4} h_K^m$. Consequently, the assertion is also proved in this case. \square

Remark 9. The proof given above extends to shape-regular quadrilateral elements (see Example 3), where $F_K(\cdot) \in (Q_2)^2$ when $m \leq 1$. In this case, the relations (34), (36), (37), and (38) hold as well. In the same way, we can treat the Clément operator for hexahedral elements K with $F_K(\cdot) \in (Q_3)^3$. The Scott–Zhang operator can also be treated if all faces are planar.

For elements with curved faces, one can use a projection operator Π_i on a reference configuration $\hat{\omega}$ (see e.g. Bernardi, 1989).

Remark 10. For error estimates of the quasi-interpolants on anisotropic meshes, we refer to Apel (1999b). The main results are the following:

1. In a suitable coordinate system (compare Remark 8), an anisotropic version of Theorem 2 holds for $m = 0$. This is obtained by a proper scaling.
2. An example shows that both quasi-interpolation operators are not suited for deriving anisotropic error estimates in the sense of (32) if $m \geq 1$.
3. Modifications of the Scott–Zhang operators have been suggested such that error estimates of type (32) can be obtained under certain assumptions on the domain ('tensor-product structure').

7 EXAMPLE FOR A GLOBAL INTERPOLATION ERROR ESTIMATE

The effectivity of numerical methods for differential and integral equations depends on the choice of the mesh. Since singularities due to the geometry of the domain are known a priori, it is advantageous to adapt the finite element mesh \mathcal{T} to these singularities. In this section, we define such meshes for a class of singularities and estimate the global interpolation error in the (broken) norm of the Sobolev space $W^{m,p}(\Omega)$. Such an error estimate is used in the estimation of the discretization error of various finite element methods. Note that this generality of the norm includes estimates in $L^2(\Omega)$, $L^\infty(\Omega)$, and $W^{1,2}(\Omega)$.

Assumption 2. Let $\Omega \subset \mathbb{R}^2$ be a two-dimensional polygonal domain with corners c_j , $j = 1, \dots, J$. The solution u of an elliptic boundary value problem has, in general, singularities near the corners c_j , that is, the solution can be represented by

$$u = u_0 + \sum_{j=1}^J u_j$$

with a regular part

$$u_0 \in W^{\ell,p}(\Omega) \quad (40)$$

and singular parts (corner singularities) u_j satisfying

$$|D^\alpha u_j| \leq C r_j^{|\alpha| - \mu_j} \quad \forall \alpha: |\alpha| \leq \ell \quad (41)$$

where $r_j = r_j(x) := \text{dist}(x, c_j)$. The integer ℓ and the real numbers p and λ_j , $j = 1, \dots, J$, are defined by the data.

This assumption is realistic for a large class of elliptic problems, including those for the Poisson equation, the Lamé system, and the biharmonic equation. The numbers λ_j depend on the geometry of Ω (in particular on the internal angles at c_j), the differential operator, and the boundary conditions. For problems with mixed boundary conditions, there are, in general, further singular terms, but since they can also be characterized by (41), this poses no extra difficulty for the forthcoming analysis.

We remark that there can be terms that are best described by

$$|D^\alpha u_j| \leq C r_j^{|\alpha| - \mu_j} |\ln r_j|^{\beta_j}$$

These terms can be treated either by a slight modification of the forthcoming analysis or by decreasing the exponent in (41) slightly. Note that $|\ln r_j|^{\beta_j} \leq C r_j^{-\epsilon}$ for all $\epsilon > 0$.

Assumption 3. Let \mathcal{T} be a finite element mesh, which is described by parameters h and $\mu_j \in (0, 1]$, $j = 1, \dots, J$. We assume that the diameter h_K of the element $K \in \mathcal{T}$ relates to the distances $r_{K,j} := \text{dist}(K, c_j)$, $j = 1, \dots, J$, according to

$$h_K \leq C h^{1/\mu_j} \quad \text{if } r_{K,j} = 0 \\ h_K \leq C h r_{K,j}^{1-\mu_j} \quad \text{if } r_{K,j} > 0 \quad \forall j = 1, \dots, J \quad (42)$$

This assumption can be satisfied when isotropic elements are used, and when the elements in the neighborhoods U_j of the corners c_j satisfy

$$C_1 h^{1/\mu_j} \leq h_K \leq C_2 h^{1/\mu_j} \quad \text{if } r_{K,j} = 0 \\ C_1 h r_{K,j}^{1-\mu_j} \leq h_K \leq C_2 h r_{K,j}^{1-\mu_j} \quad \text{if } r_{K,j} > 0 \quad (43)$$

$j = 1, \dots, J$, and when $C_1 h \leq h_K \leq C_2 h$ for all other elements. The size of the neighborhoods U_j of c_j should be independent of h but small enough such that $c_i \notin U_j$ for $i \neq j$.

Let us prove that the number of elements of such meshes is of the order h^{-2} . It is sufficient to show that the number of elements $K \subset U_j$ with $c_j \notin K$ is bounded by $C h^{-2}$. By using $\int_K 1 = C_K h_K^2$ and the relations for h_K and $r_{K,j}$, we get

$$\begin{aligned} \sum_{K \subset U_j, c_j \notin K} 1 &= \sum_{K \subset U_j, c_j \notin K} C_K^{-1} h_K^{-2} \int_K 1 \\ &\leq C h^{-2} \sum_{K \subset U_j, c_j \notin K} r_{K,j}^{-2(1-\mu_j)} \int_K 1 \\ &\leq C h^{-2} \sum_{K \subset U_j, c_j \notin K} \int_K r_j^{-2(1-\mu_j)} \\ &\leq C h^{-2} \int_{U_j} r_j^{-2(1-\mu_j)} \leq C h^{-2} \end{aligned}$$

since $\mu_j > 0$.

Finally, we remark that meshes with property (43) can be created in different ways. If the neighborhood U_j is a circular sector of radius R_j , one can just move the nodes of a uniform mesh according to the coordinate transformation

$$\frac{r_j}{R_j} \mapsto \left(\frac{r_j}{R_j} \right)^{1/\mu_j}$$

(see e.g. Raugel, 1978; Oganesyan and Rukhovets, 1979; or Apel and Milde, 1996). A second possibility is to start with a uniform mesh of mesh size h and to split all elements recursively until (43) is satisfied; see Fritsch (1990) or Apel and Milde (1996).

Assumption 4. Let V_T be a finite element space corresponding to the triangulation \mathcal{T} . Let $T: W^{\ell,p} \rightarrow V_T$ (with $(I_T u)_K = I_K(u|_K)$ for all $K \in \mathcal{T}$) be the corresponding nodal interpolation operator. We assume that it permits the local interpolation error estimate

$$|u - I_K u|_{W^{m,p}(K)} \leq C h_K^{\ell-m} |u|_{W^{\ell,p}(K)} \quad (44)$$

with ℓ, p from (40), and some $m \in \{0, \dots, \ell - 1\}$.

Note that this assumption relates the regularity of u_0 to the polynomial degree. The estimate (44) is proved in Theorem 1 only if $\mathbb{P}_{\ell-1} \subset \mathbb{P}_K$. So, if the regularity is low, then a large polynomial degree does not pay; if the polynomial degree is too low, then the regularity (40)–(41) is not fully exploited.

Theorem 3. Let the function u , the mesh \mathcal{T} , and the interpolation operator I_T satisfy Assumptions 2, 3, and 4 respectively. Then the error estimate

$$\left(\sum_{K \in \mathcal{T}} |u - I_K u|_{W^{m,p}(K)}^p \right)^{1/p} \leq C h^{\ell-m} \quad (45)$$

holds if $m \leq \ell$,

$$m - \frac{2}{p} \leq \lambda_j \quad (46)$$

$$\mu_j < \frac{\lambda_j - m + \frac{2}{p}}{\ell - m} \quad \text{if } \lambda_j < \ell - \frac{2}{p} \quad (47)$$

$$\mu_j \leq 1 \quad \text{if } \lambda_j \geq \ell - \frac{2}{p} \quad (48)$$

for all $j = 1, \dots, J$.

Note that condition (46) restricts m and p such that $u_j \in W^{m,p}(\Omega)$ is ensured,

$$|u_j|_{W^{m,p}(\Omega)} = \sum_{|\alpha|=m} \int_{\Omega} |D^\alpha u_j| \leq \int_{\Omega} r^{(\lambda_j - m)p} < \infty$$

if $(\lambda_j - m)p > -2$. With this argument, we see also that $u_j \in W^{\ell,p}(\Omega)$ if $\lambda_j \geq \ell - (2/p)$, that is, the function u_j is as regular as u_0 in this case, and no refinement ($\mu_j = 1$) is necessary in U_j .

The left-hand side of (45) is formulated with this broken Sobolev norm in order to cover the case that $I_T u \notin W^{m,p}(\Omega)$. Important applications are discretization error estimates for nonconforming finite element methods. If $I_T u \in W^{m,p}(\Omega)$, then estimate (45) can be written in the form

$$|u - I_T u|_{W^{m,p}(\Omega)} \leq C h^{\ell-m}$$

Proof. Consider the neighborhood U_j of c_j with $j \in \{1, \dots, J\}$ arbitrary but fixed. By Assumption 2, we have

$$u_i \in W^{\ell,p}(U_j), \quad i = 0, \dots, J, \quad i \neq j$$

and, therefore,

$$\begin{aligned} &\left(\sum_{K \in U_j} |u_i - I_K u_i|_{W^{m,p}(K)}^p \right)^{1/p} \\ &\leq \left(\sum_{K \in U_j} C h_K^{\ell-m} |u_i|_{W^{\ell,p}(K)}^p \right)^{1/p} \\ &\leq C h^{\ell-m}, \quad i = 0, \dots, J, \quad i \neq j \quad (49) \end{aligned}$$

Since the same argument can be applied for $\Omega \setminus \bigcup_{j=1}^J U_j$, it remains to be shown that (49) holds also for $i = j$. Note that we can assume that $\lambda_j < \ell - (2/p)$, since, otherwise, $u_j \in W^{\ell,p}(\Omega)$ and no refinement is necessary.

If $c_j \in K$, we estimate the interpolant simply by $\|I_K u_j\|_{L^\infty(K)} \leq C \|u_j\|_{L^\infty(K)} \leq Ch_K^{\lambda_j}$. Using the triangle inequality, a direct computation of $|u_j|_{W^{m,p}(K)}$, and the inverse inequality for $|I_K u_j|_{W^{m,p}(K)}$ (if $m > 0$, the case $m = 0$ is even direct), we get

$$\begin{aligned} |u_j - I_K u_j|_{W^{m,p}(K)} &\leq |u_j|_{W^{m,p}(K)} + |I_K u_j|_{W^{m,p}(K)} \\ &\leq C \left(\int_K r_j^{(\lambda_j - m)p} \right)^{1/p} + Ch_K^{\lambda_j} |K|^{1/p} \|I_K u_j\|_{L^\infty(K)} \\ &\leq Ch_K^{\lambda_j - m} |K|^{1/p} \leq Ch^{(\lambda_j - m + 2/p)/\lambda_j} \leq Ch^{\ell - m} \end{aligned}$$

where we have used the inequalities $|K| \leq Ch_K^2$, (41), (42), and (47). Note that the number of elements K with $c_j \in K$ is bounded by a constant that does not depend on h . So we get

$$\left(\sum_{K \in \mathcal{T}_h, c_j \in K} |u_j - I_K u_j|_{W^{m,p}(K)}^p \right)^{1/p} \leq Ch^{\ell - m} \quad (50)$$

Consider now an element $K \in \mathcal{T}_h$ with $c_j \notin K$, that is, with $r_{K,j} > 0$. Then, $u_j \in W^{\ell,p}(K)$ and we can use the interpolation error estimate (44). So we get

$$\begin{aligned} |u_j - I_K u_j|_{W^{m,p}(K)} &\leq Ch_K^{\ell - m} |u_j|_{W^{\ell,p}(K)} \\ &\leq Ch_K^{\ell - m} \left(\int_K r_j^{(\lambda_j - \ell)p} \right)^{1/p} \\ &\leq Ch_K^{\ell - m} \left(\int_K r_j^{(\lambda_j - \ell + (\ell - m)(1 - 1/p))p} \right)^{1/p} \\ &= Ch_K^{\ell - m} \left(\int_K r_j^{(\lambda_j - m - \mu_j(\ell - m))p} \right)^{1/p} \end{aligned}$$

since $r_{j,K} \leq r_j$ in K . Hence,

$$\begin{aligned} \left(\sum_{K \in \mathcal{T}_h, c_j \notin K} |u_j - I_K u_j|_{W^{m,p}(K)}^p \right)^{1/p} \\ \leq Ch^{\ell - m} \left(\int_{U_j} r_j^{(\lambda_j - m - \mu_j(\ell - m))p} \right)^{1/p} \end{aligned}$$

The integral on the right-hand side is finite if $[\lambda_j - m - \mu_j(\ell - m)]p > -2$, which is equivalent to (47). With (49) and (50), we conclude (45). \square

Remark 11. The given proof is an improved version of a proof for a more specific function u in a paper by Fritzsche and Oswald (1988). In this paper, the authors also address the question of the optimal choice of μ_j . They obtain for $\mu_j = [\lambda_j - m + (2/p)]/[\ell - m + (2/p)]$ the equidistribution of the element-wise interpolation error in the sense $|r_j^{\lambda_j} - I_K r_j^{\lambda_j}|_{W^{m,p}(K)} \approx \text{const.}$

Remark 12. The given proof has the advantage that it needs minimal knowledge from functional analysis. A more powerful approach is to use weighted Sobolev spaces (see Remark 6 for an example). The solutions of elliptic boundary value problems are often described by analysts in terms of different versions of such spaces; see, for example, the monographs by Grisvard (1985), Kufner and Sändig (1987), Dauge (1988), Nazarov and Plamenevsky (1994), or Kozlov, Maz'ya, and Roßmann (2001). For local and global interpolation error estimates for functions of such spaces see, for example, Grisvard (1985), Apel, Sändig, and Whiteman (1996), Apel and Nicaise (1998), or Apel, Nicaise, and Schöberl (2001). The advantage is that this approach extends to the three-dimensional case, whereas Assumption 2 is too simple to cover edge singularities.

8 RELATED CHAPTERS

(See also Chapter 4, Chapter 5, Chapter 6, Chapter 9, Chapter 17 of this Volume)

REFERENCES

- Acosta G and Durán RG. Error estimates for Q_1 isoparametric elements satisfying a weak angle condition. *SIAM J. Numer. Anal.* 2000; 38:1073–1088.
- Apel Th. Anisotropic interpolation error estimates for isoparametric quadrilateral finite elements. *Computing* 1998; 60:157–174.
- Apel Th. *Anisotropic Finite Elements: Local Estimates and Applications*. Advances in Numerical Mathematics. Teubner: Stuttgart, 1999a.
- Apel Th. Interpolation of non-smooth functions on anisotropic finite element meshes. *Math. Model. Numer. Anal.* 1999b; 33:1149–1185.
- Apel Th and Dobrowolski M. Anisotropic interpolation with applications to the finite element method. *Computing* 1992; 47:277–293.
- Apel Th and Lube G. Anisotropic mesh refinement for a singularly perturbed reaction diffusion model problem. *Appl. Numer. Math.* 1998; 26:415–433.
- Apel Th and Milde F. Comparison of several mesh refinement strategies near edges. *Commun. Numer. Methods Eng.* 1996; 12:373–381. Shortened version of Preprint SPC94.15, TU Chemnitz-Zwickau, 1994.
- Apel Th and Nicaise S. Elliptic problems in domains with edges: anisotropic regularity and anisotropic finite element meshes. In *Partial Differential Equations and Functional Analysis (In Memory of Pierre Grisvard)*, Cea J, Chensais D, Geymonat G, Lions JL (eds). Birkhäuser: Boston, 1996; 18–34. Shortened version of Preprint SPC94.16, TU Chemnitz-Zwickau, 1994.
- Apel Th and Nicaise S. The finite element method with anisotropic mesh grading for elliptic problems in domains with corners and edges. *Math. Methods Appl. Sci.* 1998; 21:519–549.
- Apel Th, Nicaise S and Schöberl J. Crouzeix-Raviart type finite elements on anisotropic meshes. *Numer. Math.* 2001; 89:193–223.
- Apel Th, Sändig A-M and Whiteman J. Graded mesh refinement and error estimates for finite element solutions of elliptic boundary value problems in non-smooth domains. *Math. Methods Appl. Sci.* 1996; 19:63–85.
- Arnold DN, Boffi D and Falk RS. Approximation by quadrilateral finite elements. *Math. Comput.* 2002; 239:909–922.
- Arunkiranthar K and Reddy BD. Some geometrical results and estimates for quadrilateral finite elements. *Comput. Methods Appl. Mech. Engrg.* 1995; 122:307–314.
- Babuska I and Aziz AK. On the angle condition in the finite element method. *SIAM J. Numer. Anal.* 1976; 13:214–226.
- Bernardi C. Optimal finite-element interpolation on curved domains. *SIAM J. Numer. Anal.* 1989; 26:1212–1240.
- Bernardi C and Girault V. A local regularization operator for triangular and quadrilateral finite elements. *SIAM J. Numer. Anal.* 1998; 35:1893–1916.
- Braess D. *Finite Elemente*. Springer: Berlin, 1997.
- Bramble JH and Hilbert SR. Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolation. *SIAM J. Numer. Anal.* 1970; 7:112–124.
- Bramble JH and Hilbert SR. Bounds for a class of linear functionals with applications to Hermite interpolation. *Numer. Math.* 1971; 16:362–369.
- Brenner SC and Scott LR. *The Mathematical Theory of Finite Element Methods*. Springer: New York, 1994.
- Carstensen C. Quasi-interpolation and a posteriori error analysis in finite element methods. *Math. Model. Numer. Anal.* 1999; 33:1187–1202.
- Ciarlet P and Raviart P-A. General Lagrange and Hermite interpolation in R^d with application to finite elements. *Arch. Ration. Mech. Anal.* 1972a; 46:177–199.
- Ciarlet P and Raviart P-A. Interpolation theory over curved elements, with applications to the finite element method. *Comput. Methods Appl. Mech. Engrg.* 1972b; 1:217–249.
- Ciarlet PG. *The Finite Element Method for Elliptic Problems*. North Holland: Amsterdam, 1972b. Reprinted by SIAM: Philadelphia, 2002.
- Ciarlet PG. Basic error estimates for elliptic problems. In *Finite Element Methods (Part I)*, vol. II of *Handbook of Numerical Analysis*, Ciarlet PG, Lions JL (eds). North Holland: Amsterdam, 1991; 17–351.
- Clément P. Approximation by finite element functions using local regularization. *RAIRO Anal. Numer.* 1975; 2:77–84.
- Dauge M. *Elliptic Boundary Value Problems on Corner Domains – Smoothness and Asymptotics of Solutions*, vol. 1341 of *Lecture Notes in Mathematics*. Springer: Berlin, 1988.
- Deny J and Lions J-L. Les espaces du type de Beppo Levi. *Ann. Inst. Fourier* 1953/54; 5:305–370.
- Dobrowolski M. *Finite Elemente*. Lecture Notes, Universität Würzburg: Würzburg, 1998.
- Dupont T and Scott R. Polynomial approximation of functions in Sobolev spaces. *Math. Comp.* 1980; 34:441–463.
- Formaggia L and Perotto S. New anisotropic a priori estimates. *Numer. Math.* 2001; 89:641–667.
- Fritzsche R. *Optimale Finite-Elemente-Approximationen für Funktionen mit Singularitäten*. PhD thesis, TU Dresden, 1990.
- Fritzsche R and Oswald P. Zur optimalen Gitterwahl bei Finite Elemente Approximationen. *Wissenschaftliche Zeitschrift TU Dresden* 1988; 37(3):155–158.
- Girault V and Raviart P-A. *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, vol. 5 of *Springer Series in Computational Mathematics*. Springer: Berlin, 1986.
- Girault V and Scott LR. Hermite interpolation of nonsmooth functions preserving boundary conditions. *Math. Comp.* 2002; 71:1043–1074.
- Gregory JA. Error bounds for linear interpolation in triangles. In *The Mathematics of Finite Elements and Applications II*, Whiteman JR (ed.). Academic Press: London, 1975; 163–170.
- Grisvard P. *Elliptic Problems in Nonsmooth Domains*, vol. 24 of *Monographs and Studies in Mathematics*. Pitman: Boston-London-Melbourne, 1985.
- Hiptmair R. Finite elements in computational electromagnetism. *Acta Numer.* 2002; 11:237–339.
- Höllig K. *Finite Element Methods with B-Splines*, vol. 26 of *Frontiers in Applied Mathematics*. SIAM: Philadelphia, 2003.
- Hughes TJR. *The Finite Element Method. Linear Static and Dynamic Finite Element Analysis*. Prentice Hall: Englewood Cliffs, 1987.
- Jamet P. Estimations d'erreur pour des éléments finis droits presque dégénérés. *RAIRO Anal. Numér.* 1976; 10:43–61.
- Jamet P. Estimation of the interpolation error for quadrilateral finite elements which can degenerate into triangles. *SIAM J. Numer. Anal.* 1977; 14:925–930.
- Kornhuber R and Ritzsch R. On adaptive grid refinement in the presence of internal and boundary layers. *IMPACT Comput. Sci. Engrg.* 1990; 2:40–72.
- Kozlov VA, Maz'ya VG and Romann J. *Spectral Problems Associated with Corner Singularities of Solutions to Elliptic Equations*. American Mathematical Society: Providence, 2001.
- Kufner A and Sändig A-M. *Some Applications of Weighted Sobolev Spaces*. Teubner: Leipzig, 1987.
- Melenk JM. *hp-Interpolation of Non-Smooth Functions*. Preprint N103050, Isaac Newton Institute for Mathematical Sciences: Cambridge, 2003.
- Ming P and Shi Z. Quadrilateral mesh. *Chin. Ann. Math., Ser. B* 2002a; 23:235–252.
- Ming P and Shi Z. Quadrilateral mesh revisited. *Comput. Methods Appl. Mech. Engrg.* 2002b; 191:5671–5682.

- Nazarov SA and Plamenevsky BA. *Elliptic Problems in Domains with Piecewise Smooth Boundary*, vol. 13 of *de Gruyter Expositions in Mathematics*. Walter de Gruyter: Berlin, 1994.
- Oden JT, Denkovicz L, Westermann TA and Rachowicz W. Toward a universal h - p adaptive finite element strategy. Part 2. A posteriori error estimates. *Comput. Methods Appl. Mech. Eng.* 1989; 77:113–180.
- Oganesyan LA and Rukhovets LA. *Variational-Difference Methods for the Solution of Elliptic Equations*. Izd. Akad. Nauk Armyanskoi SSR: Yerevan, 1979 In Russian.
- Oswald P. *Multilevel Finite Element Approximation: Theory and Applications*. Teubner: Stuttgart, 1994.
- Peraire J, Vahdati M, Morgan K and Zienkiewicz OC. Adaptive remeshing for compressible flow computation. *J. Comput. Phys.* 1987; 72:449–466.
- Raugel G. Résolution numérique par une méthode d'éléments finis du problème de Dirichlet pour le Laplacien dans un polygone. *C. R. Acad. Sci. Paris, Sér. A* 1978; 286(18): A791–A794.
- Roos H-G, Stynes M and Tobiska L. *Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion and Flow Problems*. Springer: Berlin, 1996.
- Schöberl J. *Commuting Quasi-Interpolation Operators for Mixed Finite Elements*. Preprint ISC-01-10-MATH, Texas A & M University: College Station, TX, 2001.
- Schwab C. *p - and hp -Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics*. Numerical Mathematics and Scientific Computation, Clarendon Press: Oxford, 1998.
- Scott LR and Zhang S. Finite element interpolation of non-smooth functions satisfying boundary conditions. *Math. Comp.* 1990; 54:483–493.
- Syngae JL. *The Hypercircle In Mathematical Physics*. Cambridge University Press: Cambridge, 1957.
- Verfürth R. Error estimates for some quasi-interpolation operators. *Math. Model. Numer. Anal.* 1999a; 33:695–713.
- Verfürth R. A note on polynomial approximation in Sobolev spaces. *Math. Model. Numer. Anal.* 1999b; 33:715–719.
- Zhou G and Rannacher R. Mesh orientation and anisotropic refinement in the streamline diffusion method. In *Finite Element Methods: Fifty Years of the Courant Element*, vol. 164 of *Lecture Notes in Pure and Applied Mathematics*, Křížek M, Neittaanmäki P, Stenberg R (eds). Marcel Dekker: New York, 1993; 491–500. Also published as Preprint 93-57, Universität Heidelberg, FWR, SFB 359, 1993.
- Zienkiewicz OC and Wu J. Automatic directional refinement in adaptive analysis of compressible flows. *Int. J. Numer. Methods Eng.* 1994; 37:2189–2210.
- Zlámal M. On the finite element method. *Numer. Math.* 1968; 12:394–409.

Chapter 4

Finite Element Methods

Susanne C. Brenner¹ and Carsten Carstensen²

¹University of South Carolina, Columbia, SC, USA

²Humboldt-Universität zu Berlin, Berlin, Germany

| | |
|---|-----|
| 1 Introduction | 73 |
| 2 Ritz–Galerkin Methods for Linear Elliptic Boundary Value Problems | 74 |
| 3 Finite Element Spaces | 77 |
| 4 A Priori Error Estimates for Finite Element Methods | 82 |
| 5 A Posteriori Error Estimates and Analysis | 85 |
| 6 Local Mesh Refinement | 98 |
| 7 Other Aspects | 104 |
| Acknowledgments | 114 |
| References | 114 |

1 INTRODUCTION

The finite element method is one of the most widely used techniques in computational mechanics. The mathematical origin of the method can be traced to a paper by Courant (1943). We refer the readers to the articles by Babuška (1994) and Oden (1991) for the history of the finite element method. In this chapter, we give a concise account of the h -version of the finite element method for elliptic boundary value problems in the displacement formulation, and refer the readers to the theory of Chapter 5 and Chapter 9 of this Volume.

This chapter is organized as follows. The finite element method for elliptic boundary value problems is based on the

Ritz–Galerkin approach, which is discussed in Section 2. The construction of finite element spaces and the a priori error estimates for finite element methods are presented in Sections 3 and 4. The a posteriori error estimates for finite element methods and their applications to adaptive local mesh refinements are discussed in Sections 5 and 6. For the ease of presentation, the contents of Sections 3 and 4 are restricted to symmetric problems on polyhedral domains using conforming finite elements. The extension of these results to more general situations is outlined in Section 7.

For the classical material in Sections 3, 4, and 7, we are content with highlighting the important results and pointing to the key literature. We also concentrate on basic theoretical results and refer the readers to other chapters in this encyclopedia for complications that may arise in applications. For the recent development of a posteriori error estimates and adaptive local mesh refinements in Sections 5 and 6, we try to provide a more comprehensive treatment. Owing to space limitations many significant topics and references are inevitably absent. For in-depth discussions of many of the topics covered in this chapter (and the ones that we do not touch upon), we refer the readers to the following survey articles and books (which are listed in alphabetical order) and the references therein (Ainsworth and Oden, 2000; Apel, 1999; Aziz, 1972; Babuška and Aziz, 1972; Babuška and Strouboulis, 2001; Bangerth and Rannacher, 2003; Bathe, 1996; Becker, Carey and Oden, 1981; Becker and Rannacher, 2001; Braess, 2001; Brenner and Scott, 2002; Ciarlet, 1978, 1991; Eriksson *et al.*, 1995; Hughes, 2000; Oden and Reddy, 1976; Schatz, Thomée and Wendland, 1990; Strang and Fix, 1973; Szabó and Babuška, 1991; Verfürth, 1996; Wahbin, 1991, 1995; Zienkiewicz and Taylor, 2000).

Encyclopedia of Computational Mechanics, Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes. Volume 1: *Fundamentals*. © 2004 John Wiley & Sons, Ltd. ISBN: 0-470-84699-2.

2 RITZ-GALERKIN METHODS FOR LINEAR ELLIPTIC BOUNDARY VALUE PROBLEMS

In this section, we set up the basic mathematical framework for the analysis of Ritz-Galerkin methods for linear elliptic boundary value problems. We will concentrate on symmetric problems. Nonsymmetric elliptic boundary value problems will be discussed in Section 7.1.

2.1 Weak problems

Let Ω be a bounded connected open subset of the Euclidean space \mathbb{R}^d with a piecewise smooth boundary. For a positive integer k , the Sobolev space $H^k(\Omega)$ is the space of square integrable functions whose weak derivatives up to order k are also square integrable, with the norm

$$\|v\|_{H^k(\Omega)} = \left(\sum_{|\alpha| \leq k} \left\| \frac{\partial^\alpha v}{\partial x^\alpha} \right\|_{L_2(\Omega)}^2 \right)^{1/2}$$

The seminorm $(\sum_{|\alpha|=k} \|\partial^\alpha v / \partial x^\alpha\|_{L_2(\Omega)}^2)^{1/2}$ will be denoted by $|v|_{H^k(\Omega)}$. We refer the readers to Nečas (1967), Adams (1995), Triebel (1978), Grisvard (1985), and Wloka (1987) for the properties of the Sobolev spaces. Here we just point out that $\|\cdot\|_{H^k(\Omega)}$ is a norm induced by an inner product and $H^k(\Omega)$ is complete under this norm, that is, $H^k(\Omega)$ is a Hilbert space. (We assume that the readers are familiar with normed and Hilbert spaces.)

Using the Sobolev spaces we can represent a large class of symmetric elliptic boundary value problems of order $2m$ in the following abstract weak form: Find $u \in V$, a closed subspace of a Sobolev space $H^m(\Omega)$, such that

$$a(u, v) = F(v) \quad \forall v \in V \quad (1)$$

where $F: V \rightarrow \mathbb{R}$ is a bounded linear functional on V and $a(\cdot, \cdot)$ is a symmetric bilinear form that is bounded and V -elliptic, that is,

$$|a(v_1, v_2)| \leq C_1 \|v_1\|_{H^m(\Omega)} \|v_2\|_{H^m(\Omega)} \quad \forall v_1, v_2 \in V \quad (2)$$

$$a(v, v) \geq C_2 \|v\|_{H^m(\Omega)}^2 \quad \forall v \in V \quad (3)$$

Remark 1. We use C , with or without subscript, to represent a generic positive constant that can take different values at different occurrences.

Remark 2. Equation (1) is the Euler-Lagrange equation for the variational problem of finding the minimum of the functional $v \mapsto \frac{1}{2}a(v, v) - F(v)$ on the space V . In mechanics, this functional often represents an energy and its minimization follows from the Dirichlet principle. Furthermore, the corresponding Euler-Lagrange equations (also called first variation) (1) often represent the principle of virtual work.

It follows from conditions (2) and (3) that $a(\cdot, \cdot)$ defines an inner product on V which is equivalent to the inner product of the Sobolev space $H^m(\Omega)$. Therefore the existence and uniqueness of the solution of (1) follow immediately from (2), (3), and the Riesz Representation Theorem (Yosida, 1995; Reddy, 1986; Oden and Demkowicz, 1996).

The following are typical examples from computational mechanics.

Example 1. Let $a(\cdot, \cdot)$ be defined by

$$a(v_1, v_2) = \int_{\Omega} \nabla v_1 \cdot \nabla v_2 \, dx \quad (4)$$

For $f \in L_2(\Omega)$, the weak form of the Poisson problem

$$-\Delta u = f \quad \text{on } \Omega, \quad u = 0 \quad \text{on } \Gamma, \quad \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega \setminus \Gamma \quad (5)$$

where Γ is a subset of $\partial\Omega$ with a positive $(d-1)$ -dimensional measure, is given by (1) with $V = \{v \in H^1(\Omega); v|_{\Gamma} = 0\}$ and

$$F(v) = \int_{\Omega} f v \, dx = (f, v)_{L_2(\Omega)} \quad (6)$$

For the pure Neumann problem where $\Gamma = \emptyset$, since the gradient vector vanishes for constant functions, an appropriate function space for the weak problem is $V = \{v \in H^1(\Omega); (v, 1)_{L_2(\Omega)} = 0\}$.

The boundedness of F and $a(\cdot, \cdot)$ is obvious and the coercivity of $a(\cdot, \cdot)$ follows from the Poincaré-Friedrichs inequalities (Nečas, 1967):

$$\|v\|_{L_2(\Omega)} \leq C \left(\|v\|_{H^1(\Omega)} + \left| \int_{\Gamma} v \, ds \right| \right) \quad \forall v \in H^1(\Omega) \quad (7)$$

$$\|v\|_{L_2(\Omega)} \leq C \left(\|v\|_{H^1(\Omega)} + \left| \int_{\Omega} v \, dx \right| \right) \quad \forall v \in H^1(\Omega) \quad (8)$$

Example 2. Let $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) and $v \in [H^1(\Omega)]^d$ be the displacement of an elastic body. The strain tensor $\epsilon(v)$ is given by the $d \times d$ matrix with components

$$\epsilon_{ij}(v) = \frac{1}{2} \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) \quad (9)$$

and the stress tensor $\sigma(v)$ is the $d \times d$ matrix defined by

$$\sigma(v) = 2\mu \epsilon(v) + \lambda (\operatorname{div} v) \mathbb{I} \quad (10)$$

where \mathbb{I} is the $d \times d$ identity matrix and $\mu > 0$ and $\lambda > 0$ are the Lamé constants.

Let the bilinear form $a(\cdot, \cdot)$ be defined by

$$a(v_1, v_2) = \int_{\Omega} \sum_{i,j=1}^d \sigma_{ij}(v_1) \epsilon_{ij}(v_2) \, dx = \int_{\Omega} \sigma(v_1) : \epsilon(v_2) \, dx \quad (11)$$

For $f \in [L_2(\Omega)]^d$, the weak form of the linear elasticity problem (Charlet, 1988)

$$\begin{aligned} \operatorname{div}[\sigma(u)] &= f & \text{on } \Omega, & \quad u = 0 & \text{on } \Gamma \\ [\sigma(u)]n &= 0 & \text{on } \partial\Omega \setminus \Gamma \end{aligned} \quad (12)$$

where Γ is a subset of $\partial\Omega$ with a positive $(d-1)$ -dimensional measure, is given by (1) with $V = \{v \in [H^1(\Omega)]^d; v|_{\Gamma} = 0\}$ and

$$F(v) = \int_{\Omega} f \cdot v \, dx = (f, v)_{L_2(\Omega)} \quad (13)$$

For the pure traction problem where $\Gamma = \emptyset$, the strain tensor vanishes for all infinitesimal rigid motions, i.e., displacement fields of the form $m = a + \rho x$, where $a \in \mathbb{R}^d$, ρ is a $d \times d$ antisymmetric matrix and $x = (x_1, \dots, x_d)^T$ is the position vector. In this case an appropriate function space for the weak problem is $V = \{v \in [H^1(\Omega)]^d; \int_{\Omega} \nabla \times v \, dx = 0 = \int_{\Omega} v \, dx\}$.

The boundedness of F and $a(\cdot, \cdot)$ is obvious and the coercivity of $a(\cdot, \cdot)$ follows from Korn's inequalities (Friedrichs, 1947; Duvaut and Lions, 1976; Nitsche, 1981) (see Chapter 2, Volume 2):

$$\|v\|_{H^1(\Omega)} \leq C \left(\|v\|_{L_2(\Omega)} + \left| \int_{\Gamma} v \, ds \right| \right) \quad \forall v \in [H^1(\Omega)]^d, \quad (14)$$

$$\|v\|_{H^1(\Omega)} \leq C \left(\|v\|_{L_2(\Omega)} + \left| \int_{\Omega} \nabla \times v \, dx \right| + \left| \int_{\Omega} v \, dx \right| \right) \quad \forall v \in [H^1(\Omega)]^d \quad (15)$$

Example 3. Let Ω be a domain in \mathbb{R}^2 and the bilinear form $a(\cdot, \cdot)$ be defined by

$$\begin{aligned} a(v_1, v_2) &= \int_{\Omega} \left[\Delta v_1 \Delta v_2 + (1 - \sigma) \right. \\ &\quad \times \left. \left(2 \frac{\partial^2 v_1}{\partial x_1 \partial x_2} \frac{\partial^2 v_2}{\partial x_1 \partial x_2} - \frac{\partial^2 v_1}{\partial x_1^2} \frac{\partial^2 v_2}{\partial x_2^2} - \frac{\partial^2 v_1}{\partial x_2^2} \frac{\partial^2 v_2}{\partial x_1^2} \right) \right] dx \end{aligned} \quad (16)$$

where $\sigma \in (0, 1/2)$ is the Poisson ratio.

For $f \in L_2(\Omega)$, the weak form of the clamped plate bending problem (Charlet, 1997)

$$\Delta^2 u = f \quad \text{on } \Omega, \quad u = \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega \quad (17)$$

is given by (1), where $V = \{v \in H^2(\Omega); v = \partial v / \partial n = 0 \text{ on } \partial\Omega\} = H_0^2(\Omega)$ and F is defined by (6). For the simply supported plate bending problem, the function space V is $\{v \in H^2(\Omega); v = 0 \text{ on } \partial\Omega\} = H^2(\Omega) \cap H_0^1(\Omega)$.

For these problems, the coercivity of $a(\cdot, \cdot)$ is a consequence of the following Poincaré-Friedrichs inequality (Nečas, 1967):

$$\|v\|_{H^1(\Omega)} \leq C \|v\|_{H^2(\Omega)} \quad \forall v \in H^2(\Omega) \cap H_0^1(\Omega) \quad (18)$$

Remark 3. The weak formulation of boundary value problems for beams and shells can be found in Chapter 8, this Volume and Chapter 3, Volume 2.

2.2 Ritz-Galerkin methods

In the Ritz-Galerkin approach for (1), a discrete problem is formulated as follows.

Find $\tilde{u} \in \tilde{V}$ such that

$$a(\tilde{u}, \tilde{v}) = F(\tilde{v}) \quad \forall \tilde{v} \in \tilde{V} \quad (19)$$

where \tilde{V} , the space of trial/test functions, is a finite-dimensional subspace of V .

The orthogonality relation

$$a(u - \tilde{u}, \tilde{v}) = 0 \quad \forall \tilde{v} \in \tilde{V} \quad (20)$$

follows by subtracting (19) from (1), and hence

$$\|u - \tilde{u}\|_a = \inf_{\tilde{v} \in \tilde{V}} \|u - \tilde{v}\|_a \quad (21)$$

where $\|\cdot\|_a = (a(\cdot, \cdot))^{1/2}$. Furthermore, (2), (3), and (21) imply that

$$\|u - \tilde{u}\|_{H^m(\Omega)} \leq \left(\frac{C_1}{C_2} \right)^{1/2} \inf_{\tilde{v} \in \tilde{V}} \|u - \tilde{v}\|_{H^m(\Omega)} \quad (22)$$

that is, the error for the approximate solution \tilde{u} is quasi-optimal in the norm of the Sobolev space underlying the weak problem.

The abstract estimate (22), called Cea's lemma, reduces the error estimate for the Ritz-Galerkin method to a problem in approximation theory, namely, to the determination of the magnitude of the error of the best approximation of

u by a member of \tilde{V} . The solution of this problem depends on the regularity (smoothness) of u and the nature of the space \tilde{V} .

One can also measure $u - \tilde{u}$ in other norms. For example, an estimate of $\|u - \tilde{u}\|_{L_2(\Omega)}$ can be obtained by the Aubin–Nitsche duality technique as follows. Let $w \in V$ be the solution of the weak problem

$$a(v, w) = \int_{\Omega} (u - \tilde{u})v \, dx \quad \forall v \in V \quad (23)$$

Then we have, from (20), (23), and the Cauchy–Schwarz inequality,

$$\begin{aligned} \|u - \tilde{u}\|_{L_2(\Omega)}^2 &= a(u - \tilde{u}, w) = a(u - \tilde{u}, w - \tilde{v}) \\ &\leq C_2 \|u - \tilde{u}\|_{H^m(\Omega)} \|w - \tilde{v}\|_{H^m(\Omega)} \quad \forall \tilde{v} \in \tilde{V} \end{aligned}$$

which implies that

$$\|u - \tilde{u}\|_{L_2(\Omega)} \leq C_2 \left(\inf_{\tilde{v} \in \tilde{V}} \frac{\|u - \tilde{v}\|_{H^m(\Omega)}}{\|u - \tilde{u}\|_{L_2(\Omega)}} \right) \|u - \tilde{u}\|_{H^m(\Omega)} \quad (24)$$

In general, since w can be approximated by members of \tilde{V} to high accuracy, the term inside the bracket on the right-hand side of (24) is small, which shows that the L_2 error is much smaller than the H^m error.

The estimates (22) and (24) provide the basic a priori error estimates for the Ritz–Galerkin method in an abstract setting.

On the other hand, the error of the Ritz–Galerkin method can also be estimated in an a posteriori fashion. Let the computable linear functional (the residual of the approximate solution \tilde{u}) $R: V \rightarrow \mathbb{R}$ be defined by

$$R(v) = a(u - \tilde{u}, v) = F(v) - a(\tilde{u}, v) \quad (25)$$

The global a posteriori error estimate

$$\|u - \tilde{u}\|_{H^m(\Omega)} \leq \frac{1}{C_2} \sup_{v \in \tilde{V}} \frac{|R(v)|}{\|v\|_{H^m(\Omega)}} \quad (26)$$

then follows from (3) and (25).

Let D be a subdomain of Ω and $H_0^m(D)$ be the subspace of V whose members vanish identically outside D . It follows from (25) and the local version of (2) that we also have a local a posteriori error estimate:

$$\|u - \tilde{u}\|_{H^m(D)} \geq \frac{1}{C_1} \sup_{v \in H_0^m(D)} \frac{|R(v)|}{\|v\|_{H^m(D)}} \quad (27)$$

The equivalence of the error norm with the dual norm of the residual will be the point of departure in Section 5.1.2 (cf. (70)).

2.3 Elliptic regularity

As mentioned above, the magnitude of the error of a Ritz–Galerkin method for an elliptic boundary value problem depends on the regularity of the solution. Here we give a brief description of elliptic regularity for the examples in Section 2.1.

If the boundary $\partial\Omega$ is smooth and the homogeneous boundary conditions are also smooth (i.e. the Dirichlet and Neumann boundary condition in (5) and the displacement and traction boundary conditions in (12) are defined on disjoint components of $\partial\Omega$), then the solution of the elliptic boundary value problems in Section 2.1 obey the classical *Shift Theorem* (Agmon, 1965; Nečas, 1967; Gilbarg and Trudinger, 1983; Wloka, 1987). In other words, if the right-hand side of the equation belongs to the Sobolev space $H^k(\Omega)$, then the solution of a $2m$ -th order elliptic boundary problem belongs to the Sobolev space $H^{2m+k}(\Omega)$.

The Shift Theorem does not hold for domains with piecewise smooth boundary in general. For example, let Ω be the L -shaped domain depicted in Figure 1 and

$$u(x) = \phi(r) r^{2/3} \sin\left(\frac{2}{3}\left(\theta - \frac{\pi}{2}\right)\right) \quad (28)$$

where $r = (x_1^2 + x_2^2)^{1/2}$ and $\theta = \arctan(x_2/x_1)$ are the polar coordinates and ϕ is a smooth cut-off function that equals 1 for $0 \leq r < 1/2$ and 0 for $r > 3/4$. It is easy to check that $u \in H_0^1(\Omega)$ and $-\Delta u \in C^\infty(\bar{\Omega})$. Let D be any open neighborhood of the origin in Ω . Then $u \in H^2(\Omega \setminus \bar{D})$ but $u \notin H^2(D)$. In fact u belongs to the Besov space $B_{2,\infty}^{2/3}(D)$ (Babuška and Osborn, 1991), which implies that $u \in H^{2/3-\epsilon}(D)$ for any $\epsilon > 0$, but $u \notin H^{2/3}(D)$ (see Triebel (1978) and Grisvard (1985) for a discussion of Besov spaces and fractional order Sobolev spaces). A similar situation occurs when the types of boundary condition change abruptly, such as the Poisson problem with mixed boundary conditions depicted on the circular domain in Figure 1, where the homogeneous Dirichlet boundary condition is assumed on the upper semicircle and the homogeneous Neumann boundary condition is assumed on the lower semicircle.

Therefore (Dauge, 1988), for the second (respectively fourth) order model problems in Section 2.1, the solution in general only belongs to $H^{1+\alpha}(\Omega)$ (respectively $H^{2+\alpha}(\Omega)$) for some $\alpha \in (0, 1]$ even if the right-hand side of the equation belongs to $C^\infty(\bar{\Omega})$.

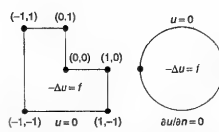


Figure 1. Singular points of two-dimensional elliptic boundary value problems.

For two-dimensional problems, the vertices of Ω and the points where the boundary condition changes type are the singular points (cf. Figure 1). Away from these singular points, the Shift Theorem is valid. The behavior of the solution near the singular points is also well understood. If the right-hand side function and its derivatives vanish to sufficiently high order at the singular points, then the Shift Theorem holds for certain weighted Sobolev spaces (Nazarov and Plamenevsky, 1994; Kozlov, Maz'ya and Rossmann, 1997, 2001). Alternatively, one can represent the solution near a singular point as a sum of a regular part and a singular part (Grisvard, 1985; Dauge, 1988; Nicaise, 1993). For a $2m$ -th order problem, the regular part of the solution belongs to the Sobolev space $H^{2m+k}(\Omega)$ if the right-hand side function belongs to $H^k(\Omega)$, and the singular part of the solution is a linear combination of special functions with less regularity, analogous to the function in (28).

The situation in three dimensions is more complicated due to the presence of edge singularities, vertex singularities, and edge-vertex singularities. The theory of three-dimensional singularities remains an active area of research.

3 FINITE ELEMENT SPACES

Finite element methods are Ritz–Galerkin methods where the finite-dimensional trial/test function spaces are constructed by piecing together polynomial functions defined on (small) parts of the domain Ω . In this section, we describe the construction and properties of finite element spaces. We will concentrate on conforming finite elements here and leave the discussion of nonconforming finite elements to Section 7.2.

3.1 The concept of a finite element

A d -dimensional finite element (Ciarlet, 1978; Brenner and Scott, 2002) is a triple $(K, \mathcal{P}_K, \mathcal{N}_K)$, where K is a closed bounded subset of \mathbb{R}^d with nonempty interior and

a piecewise smooth boundary, \mathcal{P}_K is a finite-dimensional vector space of functions defined on K and \mathcal{N}_K is a basis of the dual space \mathcal{P}_K' . The function space \mathcal{P}_K is the space of the shape functions and the elements of \mathcal{N}_K are the nodal variables (degrees of freedom).

The following are examples of two-dimensional finite elements.

Example 4 (Triangular Lagrange Elements) Let K be a triangle. The space \mathcal{P}_K of polynomials in two variables of degree $\leq n$, and let the set \mathcal{N}_K consist of evaluations of shape functions at the nodes with barycentric coordinates $\lambda_1 = i/n$, $\lambda_2 = j/n$ and $\lambda_3 = k/n$, where i, j, k are non-negative integers and $i + j + k = n$. Then $(K, \mathcal{P}_K, \mathcal{N}_K)$ is the two-dimensional P_n Lagrange finite element. The nodal variables for the P_1 , P_2 , and P_3 Lagrange elements are depicted in Figure 2, where \bullet (here and in the following examples) represents pointwise evaluation of shape functions.

Example 5 (Triangular Hermite Elements) Let K be a triangle. The cubic Hermite element is the triple $(K, \mathcal{P}_3, \mathcal{N}_K)$ where \mathcal{N}_K consists of evaluations of shape functions and their gradients at the vertices and evaluation of shape functions at the center of K . The nodal variables for the cubic Hermite element are depicted in the first figure in Figure 3, where \circ (here and in the following examples) represents pointwise evaluation of gradients of shape functions.

By removing the nodal variable at the center (cf. the second figure in Figure 3) and reducing the space of shape functions to

$$\begin{aligned} \left\{ v \in P_3; 6v(c) - 2 \sum_{i=1}^3 v(p_i) \right. \\ \left. + \sum_{i=1}^3 (\nabla v)(p_i) \cdot (p_i - c) = 0 \right\} \subset P_2 \end{aligned}$$

where p_i ($i = 1, 2, 3$) and c are the vertices and center of K respectively, we obtain the Zienkiewicz element.

The fifth degree Argyris element is the triple $(K, \mathcal{P}_5, \mathcal{N}_K)$ where \mathcal{N}_K consists of evaluations of the shape functions and their derivatives up to order two at the vertices and evaluations of the normal derivatives at the midpoints of the edges. The nodal variables for the Argyris element are depicted in the third figure in Figure 3, where \circ and \uparrow (here and



Figure 2. Lagrange elements.

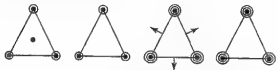


Figure 3. Cubic Hermite element, Zienkiewicz element, fifth degree Argyris element and Bell element.

in the following examples) represent pointwise evaluation of second order derivatives and the normal derivative of the shape functions, respectively.

By removing the nodal variables at the midpoints of the edges (cf. the fourth figure in Figure 3) and reducing the space of shape functions to $\{v \in P_2; (\partial v / \partial n)|_e \in P_3(e) \text{ for each edge } e\}$, we obtain the Bell element.

Example 6 (Triangular Macro Elements) Let K be a triangle that is subdivided into three subtriangles by the center of K , \mathcal{P}_K be the space of piecewise cubic polynomials with respect to this subdivision that belong to $C^1(K)$, and let the set \mathcal{N}_K consist of evaluations of the shape functions and their first-order derivatives at the vertices of K and evaluations of the normal derivatives of the shape functions at the midpoints of the edges of K . Then $(K, \mathcal{P}_K, \mathcal{N}_K)$ is the Hsieh-Clough-Tocher macro element. The nodal variables for this element are depicted in the first figure in Figure 4.

By removing the nodal variables at the midpoints of the edges (cf. the second figure in Figure 4) and reducing the space of shape functions to $\{v \in C^1(K); v \text{ is piecewise cubic and } (\partial v / \partial n)|_e \in P_3(e) \text{ for each edge } e\}$, we obtain the reduced Hsieh-Clough-Tocher macro element.

Example 7 (Rectangular Tensor Product Elements) Let K be the rectangle $[a_1, b_1] \times [a_2, b_2]$, \mathcal{P}_K be the space spanned by the monomials $x_1^i x_2^j$ for $0 \leq i, j \leq n$, and the set \mathcal{N}_K consist of evaluations of shape functions



Figure 4. Hsieh-Clough-Tocher element and reduced Hsieh-Clough-Tocher element.



Figure 5. Tensor product elements.



Figure 6. Q_4 quadrilateral elements.

at the nodes with coordinates $(a_1 + i(b_1 - a_1)/n, a_2 + j(b_2 - a_2)/n)$ for $0 \leq i, j \leq n$. Then $(K, \mathcal{P}_K, \mathcal{N}_K)$ is the two-dimensional Q_n tensor product element. The nodal variables of the Q_1 , Q_2 and Q_3 elements are depicted in Figure 5.

Example 8 (Quadrilateral Q_n Elements) Let K be a convex quadrilateral; then there exists a bilinear map $(x_1, x_2) \mapsto B(x_1, x_2) = (a_1 + b_1 x_1 + c_1 x_2 + d_1 x_1 x_2, a_2 + b_2 x_1 + c_2 x_2 + d_2 x_1 x_2)$ from the unit square S with vertices $(\pm 1, \pm 1)$ onto K . The space of shape functions is defined by $v \in \mathcal{P}_K$ if and only if $v \circ B \in Q_n$ and \mathcal{N}_K consists of pointwise evaluations of the shape functions at the nodes of K corresponding under the map B to the nodes of the Q_n tensor product element on S . The nodal variables of the Q_1 , Q_2 and Q_3 quadrilateral elements are depicted in Figure 6.

Example 9 (Other Rectangular Elements) Let K be the rectangle $[a_1, b_1] \times [c_1, d_1]$:

$$\mathcal{P}_K = \left\{ v \in Q_2; 4v(c) + \sum_{i=1}^4 v(p_i) - 2 \sum_{i=1}^4 v(m_i) = 0 \right\} \cap P_2$$

where the p_i 's are the vertices of K , the m_i 's are the midpoints of the edges of K and c is the center of K ; and \mathcal{N}_K consist of evaluations of the shape functions at the vertices and the midpoints (cf. the first figure in Figure 7). Then $(K, \mathcal{P}_K, \mathcal{N}_K)$ is the 8-node serendipity element.

If we take \mathcal{P}_K to be the space of bicubic polynomials spanned by $x_1^i x_2^j$ for $0 \leq i, j \leq 3$ and \mathcal{N}_K to be the set consisting of evaluations at the vertices of K of the shape functions, their first-order derivatives and their second-order

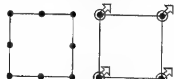


Figure 7. Serendipity and Bogner-Fox-Schmit elements.

mixed derivatives, then we have the Bogner-Fox-Schmit element. The nodal variables for this element are depicted in the second figure in Figure 7, where the tilted arrows represent pointwise evaluations of the second-order mixed derivatives of the shape functions.

Remark 4. The triangular P_n elements and the quadrilateral Q_n elements, which are suitable for second order elliptic boundary value problems, can be generalized to any dimension in a straightforward manner. The Argyris element, the Bell element, the macro elements, and the Bogner-Fox-Schmit element are suitable for fourth-order problems in two space dimensions.

3.2 Triangulations and finite element spaces

We restrict $\Omega \subset \mathbb{R}^d$ ($d = 1, 2, 3$) to be a polyhedral domain in this and the following sections. The case of curved domains will be discussed in Section 7.4.

A partition of Ω is a collection \mathcal{P} of polyhedral subdomains of Ω such that

$$\bar{\Omega} = \bigcup_{D \in \mathcal{P}} \bar{D} \quad \text{and} \quad D \cap D' = \emptyset \quad \text{if } D, D' \in \mathcal{P}, D \neq D'$$

where we use $\bar{\Omega}$ and \bar{D} to represent the closures of Ω and D .

A triangulation of Ω is a partition where the intersection of the closures of two distinct subdomains is either empty, a common vertex, a common edge or a common face. For $d = 1$, every partition is a triangulation. But the two concepts are different when $d \geq 2$. A partition that is not a triangulation is depicted in the first figure in Figure 8, where the other three figures represent triangulations. Below we will concentrate on triangulations consisting of triangles or convex quadrilaterals in two dimensions and tetrahedrons or convex hexahedrons in three dimensions.

The shape regularity of a triangle (or tetrahedron) D can be measured by the parameter

$$\gamma(D) = \frac{\text{diam } D}{\text{diameter of the largest ball in } \bar{D}} \quad (29)$$

which will be referred to as the aspect ratio of the triangle (tetrahedron). We say that a family of triangulations of

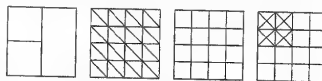


Figure 8. Partitions and triangulations.

triangles (or tetrahedrons) $\{T_i; i \in I\}$ is *regular* (or *nondegenerate*) if the aspect ratios of all the triangles (tetrahedrons) in the triangulations are bounded, that is, there exists a positive constant C such that

$$\gamma(D) \leq C \quad \text{for all } D \in T_i \text{ and } i \in I$$

The shape regularity of a convex quadrilateral (or hexahedron) D can be measured by the parameter $\gamma(D)$ defined in (29) and the parameter

$$\sigma(D) = \max \left\{ \frac{|e_1|}{|e_2|}; e_1 \text{ and } e_2 \text{ are any two edges of } D \right\} \quad (30)$$

We will refer to the number $\max(\gamma(D), \sigma(D))$ as the aspect ratio of the convex quadrilateral (hexahedron). We say that a family of triangulations of convex quadrilaterals (or hexahedrons) $\{T_i; i \in I\}$ is *regular* if the aspect ratios of all the quadrilaterals in the triangulations are bounded, that is, there exists a positive constant C such that

$$\gamma(D), \sigma(D) \leq C \quad \text{for all } D \in T_i \text{ and } i \in I$$

A family of triangulations is *quasi-uniform* if it is regular and there exists a positive constant C such that

$$h_i \leq C \text{ diam } D \quad \forall D \in T_i, i \in I \quad (31)$$

where h_i is the maximum of the diameters of the subdomains in T_i .

Remark 5. For a triangle or a tetrahedron D , a lower bound for the angles of D can lead to an upper bound for $\gamma(D)$ (and vice versa). Therefore, the *regularity* of a family of simplicial triangulations (i.e. triangulations consisting of triangles or tetrahedrons) is equivalent to the following *minimum angle condition*: There exists $\theta_0 > 0$ such that the angles of the simplexes in all the triangulations T_i are bounded below by θ_0 .

Remark 6. A family of triangulations obtained by successive uniform subdivisions of an initial triangulation is quasi-uniform. A family of triangulations generated by a *local refinement* strategy is usually regular but not quasi-uniform.

Let \mathcal{T} be a triangulation of Ω , and a finite element $(\bar{D}, \mathcal{P}_{\bar{D}}, \mathcal{N}_{\bar{D}})$ be associated with each subdomain $D \in \mathcal{T}$. We define the corresponding finite element space to be

$$FE_{\mathcal{T}} = \{v \in L_2(\Omega); v|_{\bar{D}} = v|_{\bar{D}} \in \mathcal{P}_{\bar{D}} \quad \forall D \in \mathcal{T}, \text{ and } v|_{\bar{D}}, v|_{\bar{D}'} \text{ share the same nodal values on } \bar{D} \cap \bar{D}'\} \quad (32)$$

We say that FE_T is a C^r finite element space if $FE_T \subset C^r(\bar{\Omega})$. For example, the finite element spaces constructed from the Lagrange finite elements (Example 4), the tensor product elements (Example 7), the cubic Hermite element (Example 4), the Zienkiewicz element (Example 4) and the serendipity element (Example 9) are C^0 finite element spaces, and those constructed from the quintic Argyris element (Example 5), the Bell element (Example 5), the macro elements (Example 6) and the Bogner-Fox-Schmit element (Example 9) are C^1 finite element spaces.

Note that a C^r finite element space is automatically a subspace of the Sobolev space $H^{r+1}(\Omega)$ and therefore appropriate for elliptic boundary value problems of order $2(r+1)$.

3.3 Element nodal interpolation operators and interpolation error estimates

Let $(K, \mathcal{P}_K, \mathcal{N}_K)$ be a finite element. Denote the nodal variables in \mathcal{N}_K by N_1, \dots, N_n ($n = \dim \mathcal{P}_K$) and the dual basis of \mathcal{P}_K by ϕ_1, \dots, ϕ_n , that is,

$$N_i(\phi_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Assume that $\zeta \mapsto N_i(\zeta)$ is well-defined for $\zeta \in H^s(K)$ (where s is a sufficiently large positive number), then we can define the element nodal interpolation operator $\Pi_K: H^s(K) \rightarrow \mathcal{P}_K$ by

$$\Pi_K \zeta = \sum_{j=1}^n N_j(\zeta) \phi_j \quad (33)$$

Note that (33) implies

$$\Pi_K v = v \quad \forall v \in \mathcal{P}_K \quad (34)$$

For example, by the Sobolev embedding theorem (Adams, 1995; Nečas, 1967; Wloka, 1987; Gilbarg and Trudinger, 1983), the nodal interpolation operators associated with the Lagrange finite elements (Example 4), the tensor product finite elements (Example 7), and the serendipity element (Example 9) are well-defined on $H^s(K)$ for $s > 1$ if $K \subset \mathbb{R}^2$ and for $s > 3/2$ if $K \subset \mathbb{R}^3$. On the other hand the nodal interpolation operators associated with the Zienkiewicz element (Example 5) and the macro elements (Example 6) are well-defined on $H^s(K)$ for $s > 2$, while the interpolation operators for the quintic Argyris element (Example 5), the Bell element (Example 5) or the Bogner-Fox-Schmit (Example 9) are well-defined on $H^s(K)$ for $s > 3$.

The error of the element nodal interpolation operator for a triangular (tetrahedral) or convex quadrilateral (hexagonal)

element $(K, \mathcal{P}_K, \mathcal{N}_K)$ can be controlled in terms of the shape regularity of K . Let \tilde{K} be the image of K under the scaling map

$$x \mapsto \mathcal{H}(x) = (\text{diam } K)^{-1}x \quad (35)$$

Then \tilde{K} is a domain of unit diameter and we can define a finite element $(\tilde{K}, \mathcal{P}_{\tilde{K}}, \mathcal{N}_{\tilde{K}})$ as follows: (i) $\tilde{v} \in \mathcal{P}_{\tilde{K}}$ if and only if $\tilde{v} \circ \mathcal{H} \in \mathcal{P}_K$, and (ii) $\tilde{N} \in \mathcal{N}_{\tilde{K}}$ if and only if the linear functional $\tilde{v} \mapsto N(\tilde{v} \circ \mathcal{H}^{-1})$ on $\mathcal{P}_{\tilde{K}}$ belongs to \mathcal{N}_K . It follows that the dual basis $\tilde{\phi}_1, \dots, \tilde{\phi}_n$ of $\mathcal{P}_{\tilde{K}}$ is related to the dual basis ϕ_1, \dots, ϕ_n of \mathcal{P}_K through the relation $\tilde{\phi}_i \circ \mathcal{H} = \phi_i$, and (33) implies that

$$(\Pi_K \zeta) \circ \mathcal{H}^{-1} = \Pi_{\tilde{K}}(\zeta \circ \mathcal{H}^{-1}) \quad (36)$$

for all sufficiently smooth functions ζ defined on K . Moreover, for the functions $\tilde{\zeta}$ and ζ related by $\zeta(x) = \tilde{\zeta}(\mathcal{H}(x))$, we have

$$|\tilde{\zeta}|_{H^s(\tilde{K})}^2 = (\text{diam } K)^{2s-d} |\zeta|_{H^s(K)}^2 \quad (37)$$

where d is the spatial dimension.

Assuming that $\mathcal{P}_{\tilde{K}} \supseteq \mathcal{P}_m$ (equivalently $\mathcal{P}_K \supseteq \mathcal{P}_m$), we have, by (34),

$$\begin{aligned} \|\zeta - \Pi_K \zeta\|_{H^s(K)} &= \|(\zeta - p) - \Pi_{\tilde{K}}(\zeta - p)\|_{H^s(K)} \\ &\leq 2\|\Pi_{\tilde{K}}\|_{m,s} \|\zeta - p\|_{H^s(\tilde{K})} \quad \forall p \in \mathcal{P}_m \end{aligned}$$

where $\|\Pi_{\tilde{K}}\|_{m,s}$ is the norm of the operator $\Pi_{\tilde{K}}: H^s(\tilde{K}) \rightarrow H^m(\tilde{K})$, and hence

$$\|\zeta - \Pi_K \zeta\|_{H^s(K)} \leq 2\|\Pi_{\tilde{K}}\|_{m,s} \inf_{p \in \mathcal{P}_m} \|\zeta - p\|_{H^s(\tilde{K})} \quad (38)$$

Since K is convex, the following estimate (Verfürth, 1999) holds provided m is the largest integer strictly less than s :

$$\inf_{p \in \mathcal{P}_m} \|\zeta - p\|_{H^s(K)} \leq C_{s,d} |\zeta|_{H^s(K)} \quad \forall \zeta \in H^s(K) \quad (39)$$

where the positive constant $C_{s,d}$ depends only on s and d . Combining (38) and (39) we find

$$\|\zeta - \Pi_K \zeta\|_{H^s(K)} \leq 2C_{s,d} \|\Pi_{\tilde{K}}\|_{m,s} |\zeta|_{H^s(K)} \quad \forall \zeta \in H^s(K) \quad (40)$$

We have therefore reduced the error estimate for the element nodal interpolation operator to an estimate of $\|\Pi_{\tilde{K}}\|_{m,s}$. Since $\text{diam } \tilde{K} = 1$, the norm $\|\Pi_{\tilde{K}}\|_{m,s}$ is a constant depending only on the shape of \tilde{K} (equivalently of K), if we considered s and m to be fixed for a given type of element.

For triangular elements, we can use the concept of *affine-interpolation-equivalent* elements to obtain a more concrete

description of the dependence of $\|\Pi_{\tilde{K}}\|_{m,s}$ on the shape of \tilde{K} . A d -dimensional nondegenerate affine map is a map of the form $x \mapsto Ax + b$ where A is a nonsingular $d \times d$ matrix and $b \in \mathbb{R}^d$. We say that two finite elements $(K_1, \mathcal{P}_{K_1}, \mathcal{N}_{K_1})$ and $(K_2, \mathcal{P}_{K_2}, \mathcal{N}_{K_2})$ are affine-equivalent if (i) there exists a nondegenerate affine map Φ that maps K_1 onto K_2 , (ii) $v \in \mathcal{P}_{K_1}$ if and only if $v \circ \Phi \in \mathcal{P}_{K_2}$ and (iii)

$$(\Pi_{K_1} \zeta) \circ \Phi = \Pi_{K_2}(\zeta \circ \Phi) \quad (41)$$

for all sufficiently smooth functions ζ defined on K_2 . For example, any triangular elements in one of the families (except the Bell element and the reduced Hsieh-Clough-Tocher element) described in Section 3.1 are affine-interpolation-equivalent to the corresponding element on the standard simplex \tilde{S} with vertices $(0, 0)$, $(1, 0)$ and $(0, 1)$.

Assuming $(\tilde{K}, \mathcal{P}_{\tilde{K}}, \mathcal{N}_{\tilde{K}})$ (or equivalently $(K, \mathcal{P}_K, \mathcal{N}_K)$) is affine-interpolation-equivalent to the element $(S, \mathcal{P}_S, \mathcal{N}_S)$ on the standard simplex, it follows from (41) and the chain rule that

$$\|\Pi_{\tilde{K}}\|_{m,s} \leq C \|\Pi_S\|_{m,s} \quad (42)$$

where the positive constant depends only on the Jacobian matrix of the affine map $\Phi: S \rightarrow \tilde{K}$ and thus depends only on an upper bound of the parameter $\gamma(\tilde{K})$ (cf. (29)) which is identical with $\gamma(K)$.

Combining (36), (37), (40) and (42), we find

$$\begin{aligned} \sum_{k=0}^m (\text{diam } K)^k |\zeta - \Pi_K \zeta|_{H^k(K)} &\leq C (\text{diam } K)^s |\zeta|_{H^s(K)} \\ \forall \zeta \in H^s(K) \end{aligned} \quad (43)$$

where the positive constant C depends only on s and an upper bound of the parameter $\gamma(K)$ (the aspect ratio of K), provided that (i) the element nodal interpolation operator is well-defined on $H^s(K)$, (ii) the triangular element $(K, \mathcal{P}_K, \mathcal{N}_K)$ is affine-interpolation-equivalent to a reference element $(S, \mathcal{P}_S, \mathcal{N}_S)$ on the standard simplex, (iii) $\mathcal{P} \supseteq \mathcal{P}_m$, and (iv) m is the largest integer $< s$.

For convex quadrilateral elements, we can similarly obtain a concrete description of the dependence of $\|\Pi_{\tilde{K}}\|_{m,s}$ on the shape of \tilde{K} by assuming that there is a reference element $(S, \mathcal{P}_S, \mathcal{N}_S)$ defined on the unit square S with vertices $(\pm 1, \pm 1)$ and a bilinear homeomorphism Φ from S onto \tilde{K} with the following properties: $\tilde{v} \in \mathcal{P}_{\tilde{K}}$ if and only if $v \circ \Phi \in \mathcal{P}_S$ and $(\Pi_{\tilde{K}} \zeta) \circ \Phi = \Pi_S(\zeta \circ \Phi)$ for all sufficiently smooth functions ζ defined on \tilde{K} . Note that because of (36)

this is equivalent to the existence of a bilinear homeomorphism from S onto K such that

$$v \in \mathcal{P}_K \iff v \circ \Phi \in \mathcal{P}_S \quad \text{and} \quad (\Pi_K \zeta) \circ \Phi = \Pi_S(\zeta \circ \Phi) \quad (44)$$

for all sufficiently smooth functions ζ defined on K . The estimate (42) holds again by the chain rule, where the positive constant C depends only on the Jacobian matrix of Φ and thus depends only on upper bounds for the parameters $\gamma(K)$ and $\sigma(K)$ (cf. (30)), which are identical with $\gamma(K)$ and $\sigma(K)$. We conclude that the estimate (43) also holds for convex quadrilateral elements where the positive constant C depends on upper bounds of the aspect ratio of K (equivalently an upper bound of the aspect ratio of \tilde{K}) provided condition (ii) is replaced by (44). For example, the estimate (43) is valid for the quadrilateral Q_4 element in Example 8.

Remark 7. The general estimate (40) can be refined to yield *anisotropic* error estimates for certain reference elements. For example, in two dimensions, the following estimates (Apel and Dobrowolski, 1992; Apel, 1999) hold for the P_4 Lagrange elements on the reference simplex S and the Q_4 tensor product elements on the reference square S :

$$\begin{aligned} \left| \frac{\partial}{\partial x_j} (\zeta - \Pi_S \zeta) \right|_{L_2(S)} &\leq C \left(\left| \frac{\partial^2 \zeta}{\partial x_1 \partial x_j} \right|_{L_2(S)} + \left| \frac{\partial^2 \zeta}{\partial x_2 \partial x_j} \right|_{L_2(S)} \right) \quad (45) \end{aligned}$$

for $j = 1, 2$ and for all $\zeta \in H^2(S)$. We refer the readers to Chapter 3, this Volume, for more details.

Remark 8. The analysis of the quadrilateral serendipity elements is more subtle. A detailed discussion can be found in Arnold, Boffi and Falk (2002).

Remark 9. The estimate (43) can be generalized naturally to 3-D tetrahedral P_n elements and hexahedral Q_n elements.

Remark 10. Let n be a nonnegative integer and $n < s \leq n+1$. The estimate

$$\inf_{p \in \mathcal{P}_n(\Omega)} \|\zeta - p\|_{H^s(\Omega)} \leq C_{\Omega,s} |\zeta|_{H^s(\Omega)} \quad \forall \zeta \in H^s(\Omega) \quad (46)$$

for general Ω follows from generalized Poincaré-Friedrichs inequalities (Nečas, 1967). In the case where Ω is convex, the constant $C_{\Omega,s}$ depends only on s and the dimension of Ω , but not on the shape of Ω , as indicated by the estimate (39). For nonconvex domains, the constant $C_{\Omega,s}$

does depend on the shape of Ω (Dupont and Scott, 1980, Verfürth, 1999).

Let F be a bounded linear functional on $H^1(\Omega)$ with norm $\|F\|$ such that $F(p) = 0$ for all $p \in P_n(\Omega)$. It follows from (46) that

$$|F(\zeta)| \leq \inf_{p \in P_n(\Omega)} |F(\zeta - p)| \leq \|F\| \inf_{p \in P_n(\Omega)} \|\zeta - p\|_{H^1(\Omega)} \leq (C_{\Omega,s} \|F\|) \|\zeta\|_{H^1(\Omega)} \quad (47)$$

for all $\zeta \in H^1(\Omega)$. The estimate (47), known as the Bramble–Hilbert lemma (Bramble and Hilbert, 1970), is useful for deriving various error estimates.

3.4 Some discrete estimates

The finite element spaces in Section 3.2 are designed to be subspaces of Sobolev spaces so that they can serve as the trial/test spaces for Ritz–Galerkin methods. On the other hand, since finite element spaces are constructed by piecing together finite-dimensional function spaces, there are discrete estimates valid on the finite element spaces but not the Sobolev spaces.

Let $(K, \mathcal{P}_K, \mathcal{N}_K)$ be a finite element such that $\mathcal{P}_K \subset H^k(K)$ for a nonnegative integer k . Since any seminorm on a finite-dimensional space is continuous with respect to a norm, we have, by scaling, the following inverse estimate:

$$|v|_{H^{\ell}(K)} \leq C(\text{diam } K)^{\ell-k} \|v\|_{H^k(K)} \quad \forall v \in \mathcal{P}_K, \quad 0 \leq \ell \leq k \quad (48)$$

where the positive constant C depends on the domain \tilde{K} (the image of K under the scaling map \mathcal{H} defined by (35)) and the space \mathcal{P}_K .

For finite elements whose shape functions can be pulled back to a fixed finite-dimensional function space on a reference element, the constant C depends only on the shape regularity of the element domain K and global versions of (48) can be easily derived. For example, for a quasi-uniform family $\{T_i; i \in I\}$ of simplicial or quadrilateral triangulations of a polygonal domain Ω , we have

$$|v|_{H^{\ell}(\Omega)} \leq Ch_i^{-1} \|v\|_{L_2(\Omega)} \quad \forall v \in V_i \quad \text{and} \quad i \in I \quad (49)$$

where $V_i \subset H^1(\Omega)$ is either the P_n triangular finite element space or the Q_n quadrilateral finite element space associated with T_i . Note that $V_i \subset H^1(\Omega)$ for any $s < 3/2$ and a bit more work shows that the following inverse estimate (Ben Belgacem and Brenner, 2001) also holds:

$$|v|_{H^s(\Omega)} \leq C_s h_i^{1-s} \|v\|_{H^1(\Omega)} \quad \forall v \in V_i, \quad i \in I \quad \text{and} \quad 1 \leq s < 3/2 \quad (50)$$

where the positive constant C_s can be uniformly bounded for s in a compact subset of $[1, 3/2]$.

It is well-known that in two dimensions the Sobolev space $H^1(\Omega)$ is not a subspace of $C(\bar{\Omega})$. However, the P_n triangular finite element space and the Q_n quadrilateral finite element space do belong to $C(\bar{\Omega})$ and it is possible to bound the L_∞ norm of the finite element function by its H^1 norm. Indeed, it follows from Fourier transform and extension theorems (Adams, 1995; Wloka, 1987) that, for $\epsilon > 0$,

$$\|v\|_{L_\infty(\Omega)} \leq C\epsilon^{-1/2} \|v\|_{H^{1+\epsilon}(\Omega)} \quad \forall v \in H^{1+\epsilon}(\Omega) \quad (51)$$

By taking $\epsilon = (1 + |\ln h_i|)^{-1}$ in (51) and applying (50), we arrive at the following discrete Sobolev inequality:

$$\|v\|_{L_\infty(\Omega)} \leq C(1 + |\ln h_i|)^{1/2} \|v\|_{H^1(\Omega)} \quad \forall v \in V_i \quad \text{and} \quad i \in I \quad (52)$$

where the positive constant C is independent of $i \in I$.

The discrete Sobolev inequality and the Poincaré–Friedrichs inequality (8) imply immediately the following discrete Poincaré inequality:

$$\begin{aligned} \|v\|_{L_\infty(\Omega)} &\leq \|v - \bar{v}\|_{L_\infty(\Omega)} + \|\bar{v}\|_{L_\infty(\Omega)} \leq 2\|v - \bar{v}\|_{L_\infty(\Omega)} \\ &\leq C(1 + |\ln h_i|)^{1/2} \|v - \bar{v}\|_{H^1(\Omega)} \\ &\leq C(1 + |\ln h_i|)^{1/2} |v|_{H^1(\Omega)} \end{aligned} \quad (53)$$

for all $v \in V_i$ that vanishes at a given point in $\bar{\Omega}$ and with mean $\bar{v} = \int_\Omega v \, dx / |\Omega|$.

Remark 11. The discrete Sobolev inequality can also be established directly using calculus and inverse estimates (Bramble, Pasciak and Schatz, 1986; Brenner and Scott, 2002), and both (52) and (53) are sharp (Brenner and Sung, 2000).

4 A PRIORI ERROR ESTIMATES FOR FINITE ELEMENT METHODS

Let T be a triangulation of Ω and a finite element $(\bar{D}, \mathcal{P}_{\bar{D}}, \mathcal{N}_{\bar{D}})$ be associated with each subdomain $D \in T$ so that the resulting finite element space FE_T (cf. (32)) is a subspace of $C^{m-1}(\bar{\Omega}) \subset H^m(\Omega)$. By imposing appropriate boundary conditions, we can obtain a subspace V_T of FE_T such that $V_T \subset V$, the subspace of $H^m(\Omega)$, where the weak problem (1) is formulated. The corresponding finite element method for (1) is:

Find $u_T \in V_T$ such that

$$a(u_T, v) = F(v) \quad \forall v \in V_T \quad (54)$$

In this section, we consider a priori estimates for the discretization error $u - u_T$. We will discuss the second-order and fourth-order cases separately. We use the letter C to denote a generic positive constant that can take different values at different appearances.

Let us also point out that the asymptotic error analysis carried out in this section is not sufficient for parameter-dependent problems (e.g. thin structures and nearly incompressible elasticity) that can experience *locking* (Babuška and Suri, 1992). We refer the readers to other chapters in this encyclopedia that are devoted to such problems for the discussion of the techniques that can overcome locking.

4.1 Second-order problems

We will devote most of our discussion to the case where $\Omega \subset \mathbb{R}^2$ and only comment briefly on the 3-D case. For preciseness, we also assume the right-hand side of the elliptic boundary value problem to be square integrable. We first consider the case where $V \subset H^1(\Omega)$ is defined by homogeneous Dirichlet boundary conditions (cf. Section 2.1) on $\Gamma \subset \partial\Omega$. Such problems can be discretized by triangular P_n elements (Example 4) or quadrilateral Q_n elements (Example 8).

Let T be a triangulation of Ω by triangles (convex quadrilaterals) and each triangle (quadrilateral) in T be equipped with the P_n ($n \geq 1$) Lagrange element (Q_n quadrilateral element). The resulting finite element space FE_T is a subspace of $C^0(\bar{\Omega}) \subset H^1(\Omega)$. We assume that Γ is the union of the edges of the triangles (quadrilaterals) in T and take $V_T = V \cap FE_T$, the subspace defined by the homogeneous Dirichlet boundary condition on Γ .

We know from the discussion in Section 2.3 that $u \in H^{1+\alpha(D)}(D)$ for each $D \in T$, where the number $\alpha(D) \in (0, 1)$ and $\alpha(D) = 1$ for D away from the singular points. Hence, the element nodal interpolation operator Π_D^1 is well-defined on u for all $D \in T$. We can therefore piece together a global nodal interpolant $\Pi_T^1 u \in V_T$ by the formula

$$(\Pi_T^1 u)|_D = \Pi_D^1(u|_D) \quad (55)$$

From the discussion in Section 3.3, we know that (43) is valid for both the triangular P_n element and the quadrilateral Q_n element. We deduce from (43) and (55) that

$$\|u - \Pi_T^1 u\|_{H^1(\Omega)}^2 \leq C \sum_{D \in T} (\text{diam } D)^{2\alpha(D)} |u|_{H^{1+\alpha(D)}(D)}^2 \quad (56)$$

where C depends only on the maximum of the aspect ratios of the element domains in T . Combining (22) and (56) we

have the a priori discretization error estimate

$$\|u - u_T\|_{H^1(\Omega)} \leq C \left(\sum_{D \in T} (\text{diam } D)^{2\alpha(D)} |u|_{H^{1+\alpha(D)}(D)}^2 \right)^{1/2} \quad (57)$$

where C depends only on the constants in (2) and (3) and the maximum of the aspect ratios of the element domains in T .

Hence, if $\{T_i; i \in I\}$ is a regular family of triangulations, and the solution u of (1) belongs to the Sobolev space $H^{1+\alpha}(\Omega)$ for some $\alpha \in (0, 1]$, then we can deduce from (57) that

$$\|u - u_T\|_{H^1(\Omega)} \leq Ch_T^\alpha |u|_{H^{1+\alpha}(\Omega)} \quad (58)$$

where $h_i = \max_{D \in T_i} \text{diam } D$ is the mesh size of T_i and C is independent of $i \in I$. Note that the solution w of (23) with \bar{u} replaced by u_T also belongs to $H^{1+\alpha}(\Omega)$ and satisfies the elliptic regularity estimate

$$\|w\|_{H^{1+\alpha}(\Omega)} \leq C \|u - u_T\|_{L_2(\Omega)}$$

Therefore, we have

$$\begin{aligned} \inf_{v \in V_T} \|w - v\|_{H^1(\Omega)} &\leq \|w - \Pi_T^1 w\|_{H^1(\Omega)} \leq Ch_T^\alpha |w|_{H^{1+\alpha}(\Omega)} \\ &\leq Ch_T^\alpha \|u - u_T\|_{L_2(\Omega)} \end{aligned} \quad (59)$$

The abstract estimate (24) with \bar{u} replaced by u_T and (59) yield the following L_2 estimate:

$$\|u - u_T\|_{L_2(\Omega)} \leq Ch_T^{2\alpha} |u|_{H^{1+\alpha}(\Omega)} \quad (60)$$

where C is also independent of $i \in I$.

Remark 12. In the case where $\alpha = 1$ (for example, when $\Gamma = \partial\Omega$ in Example 1 and Ω is convex), the estimate (58) is optimal and it is appropriate to use a quasi-uniform family of triangulations. In the case where $\alpha(D) < 1$ for D next to singular points, the estimate (57) allows the possibility of improvement by graded meshes (cf. Section 6).

Remark 13. In the derivations of (58) and (60) above for the triangular P_n elements, we have used the minimum angle condition (cf. Remark 5). In view of the anisotropic estimates (45), these estimates also hold for triangular P_n elements under the *maximum angle condition* (Babuška and Aziz, 1976; Jamet, 1976; Ženišek, 1995; Apel, 1999): there exists $\theta_0 < \pi$ such that all the angles in the family of triangulations are $\leq \theta_0$. The estimates (58) and (60) are also valid for Q_n elements on parallelograms satisfying the maximum angle condition. They can also be established for certain thin quadrilateral elements (Apel, 1999).

The 2-D results above also hold for 3-D tetrahedral P_n elements and 3-D hexagonal Q_n elements if the solution u of (1) belongs to $H^{1+\alpha}(\Omega)$ where $1/2 < \alpha \leq 1$, since the nodal interpolation operator are then well-defined by the Sobolev embedding theorem. This is the case, for example, if $\Gamma = \partial\Omega$ in Example 1. However, new interpolation operators that require less regularity are needed if $0 < \alpha \leq 3/2$. Below, we construct an interpolation operator $\Pi_D^A: H^1(\Omega) \rightarrow V_T$ using the local averaging technique of Scott and Zhang (1990).

For simplicity, we take V_T to be a tetrahedral P_1 finite element space. Therefore, we only need to specify the value of $\Pi_D^A \zeta$ at the vertices of T for a given function $\zeta \in H^1(\Omega)$. Let p be a vertex. We choose a face (or edge in 2-D) \mathcal{F} of a subdomain in T such that $p \in \mathcal{F}$. The choice of \mathcal{F} is of course not unique. But we always choose $\mathcal{F} \subset \partial\Omega$ if $p \in \partial\Omega$ so that the resulting interpolant will satisfy the appropriate Dirichlet boundary condition. Let $\{\psi_j\}_{j=1}^d \subset P_1(\mathcal{F})$ be biorthogonal to the nodal basis $\{\phi_j\}_{j=1}^d \subset P_1(\mathcal{F})$ with respect to the $L_2(\mathcal{F})$ inner product. In other words ϕ_j equals 1 at the j th vertex of \mathcal{F} and vanishes at the other vertices, and

$$\int_{\mathcal{F}} \psi_j \phi_i ds = \delta_{ij} \quad (61)$$

Suppose p corresponds to the j th vertex of \mathcal{F} . We then define

$$(\Pi_D^A \zeta)(p) = \int_{\mathcal{F}} \psi_j \zeta ds \quad (62)$$

where the integral is well-defined because of the trace theorem.

It is clear in view of (61) and (62) that $\Pi_D^A v = v$ for all $v \in FE_T$ and $\Pi_D^A \zeta = 0$ on Γ if $\zeta = 0$ on Γ . Note also that Π_D^A is not a local operator, i.e., $(\Pi_D^A \zeta)|_D$ is in general determined by $\zeta|_{S(D)}$, where $S(D)$ is the polyhedral domain formed by the subdomains in T sharing (at least) a vertex with D (cf. Figure 9 for a 2-D example). It follows that the interpolation error estimate for Π_T takes the following form:

$$\begin{aligned} \|\zeta - \Pi_D^A \zeta\|_{L_2(D)}^2 + (\text{diam } D)^2 \|\zeta - \Pi_D^A \zeta\|_{H^1(D)}^2 \\ \leq C (\text{diam } D)^{2(1+\alpha(S(D)))} \|\zeta\|_{H^{1+\alpha(S(D))}}^2 \end{aligned} \quad (63)$$

where C depends on the shape regularity of T , provided that $\zeta \in H^{1+\alpha(S(D))}(S(D))$ for some $\alpha(S(D)) \in (0, 1]$. The estimates (58) and (60) for tetrahedral P_1 elements can be derived for general $\alpha \in (0, 1]$ and regular triangulations using the estimate (63).

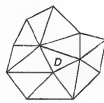


Figure 9. A two-dimensional example of $S(D)$.

Remark 14. The interpolation operator Π_D^A can be defined for general finite elements (Scott and Zhang, 1990; Girault and Scott, 2002) and anisotropic estimates can be obtained for Π_D^A for certain triangulations (Apel, 1999). There also exist interpolation operators for less regular functions (Clément, 1975; Bernardi and Girault, 1998).

Next, we consider the case where V is a closed subspace of $H^1(\Omega)$ with finite codimension $n < \infty$, as in the case of the Poisson problem with pure homogeneous Neumann boundary condition (where $n = 1$) or the elasticity problem with pure homogeneous traction boundary condition (where $n = 1$ when $d = 1$, and $n = 3(d - 1)$ when $d = 2$ or 3). The key assumption here is that there exists a bounded linear projection Q from $H^1(\Omega)$ onto an n dimensional subspace of FE_T such that $\zeta \in H^1(\Omega)$ belongs to V if and only if $Q\zeta = 0$. We can then define an interpolation operator $\tilde{\Pi}_T$ from appropriate Sobolev spaces onto V_T by

$$\tilde{\Pi}_T = (I - Q)\Pi_T$$

where Π_T is either the nodal interpolation operator Π_T^N or the Scott-Zhang averaging interpolation operator Π_T^A introduced earlier. Observe that, since the weak solution u belongs to V ,

$$u - \tilde{\Pi}_T u = u - Qu - (I - Q)\Pi_T u = (I - Q)(u - \Pi_T u)$$

and the interpolation error of $\tilde{\Pi}_T$ can be estimated in terms of the norm of $Q: H^1(\Omega) \rightarrow H^1(\Omega)$ and the interpolation error of Π_T . Therefore, the a priori discretization error estimates for Dirichlet or Dirichlet/Neumann boundary value problems also hold for this second type of (pure Neumann) boundary value problems.

For the Poisson problem with homogeneous Neumann boundary condition, we can take

$$Q\zeta = \frac{1}{|\Omega|} \int_{\Omega} \zeta dx$$

the mean of ζ over Ω . For the elasticity problem with pure homogeneous traction boundary condition, the operator Q from $[H^1(\Omega)]^d$ onto the space of infinitesimal rigid motions

is defined by

$$\begin{aligned} \int_{\Omega} Q\zeta dx &= \int_{\Omega} \zeta dx \quad \text{and} \\ \int_{\Omega} \nabla \times Q\zeta dx &= \int_{\Omega} \nabla \times \zeta dx \quad \forall \zeta \in [H^1(\Omega)]^d \end{aligned}$$

In both cases, the norm of Q is bounded by a constant C_{Ω} .

Remark 15. In the case where $f \in H^k(\Omega)$ for $k > 0$, the solution u belongs to $H^{2+k}(\Omega)$ away from the geometric or boundary data singularities and, in particular, away from $\partial\Omega$. Therefore, it is advantageous to use higher-order elements in certain parts of Ω , or even globally (with curved elements near $\partial\Omega$) if singularities are not present. In the case where $f \in L_2(\Omega)$, the error estimate (57) indicates that the order of the discretization error for the triangular P_n element or the quadrilateral Q_n element is independent of $n \geq 1$. However, the convergence of the finite element solutions to a particular solution as $h \downarrow 0$ can be improved by using higher-order elements because of the existence of *nonuniform error estimates* (Babuška and Kellogg, 1975).

4.2 Fourth-order problems

We restrict the discussion of fourth-order problems to the two-dimensional plate bending problem of Example 3.

Let T be a triangulation of Ω by triangles and each triangle in T be equipped with the Hsieh-Clough-Tocher macro element (cf. Example 6). The finite element space FE_T defined by (32) is a subspace of $C^1(\bar{\Omega}) \subset H^2(\Omega)$. We take V_T to be $V \cap FE_T$, where $V = H_0^2(\Omega)$ for the clamped plate and $V = H_0^2(\Omega) \cap H^2(\Omega)$ for the simply supported plate.

The solution u of the plate-bending problem belongs to $H^{2+\alpha(D)}(D)$ for each $D \in T$, where $\alpha(D) \in (0, 2]$ and $\alpha(D) = 2$ for D away from the corners of Ω . The elemental nodal interpolation operator Π_D^N is well-defined on u for all $D \in T$. We can therefore define a global nodal interpolation operator Π_T^N by the formula (55). Since the Hsieh-Clough-Tocher macro element is affine-interpolation-equivalent to the reference element on the standard simplex, we deduce from (55) and (43) that

$$\|u - \Pi_T^N u\|_{H^2(\Omega)}^2 \leq C \sum_{D \in T} (\text{diam } D)^{2\alpha(D)} \|u\|_{H^{2+\alpha(D)}(D)}^2 \quad (64)$$

where C depends only on the maximum of the aspect ratios of the triangles in T (or equivalently the minimum angle of

T). From (22) and (64), we have

$$\|u - u_T\|_{H^2(\Omega)} \leq C \left(\sum_{D \in T} (\text{diam } D)^{2\alpha(D)} \|u\|_{H^{2+\alpha(D)}(D)}^2 \right)^{1/2} \quad (65)$$

where C depends only on the constants in (2) and (3) and the minimum angle of T .

Hence, if $\{T_i; i \in I\}$ is a regular family of triangulations, and the solution u of the plate bending problem belongs to the Sobolev space $H^{2+\alpha}(\Omega)$ for some $\alpha \in (0, 2]$, we can deduce from (65) that

$$\|u - u_T\|_{H^2(\Omega)} \leq Ch_i^2 \|u\|_{H^{2+\alpha}(\Omega)} \quad (66)$$

where $h_i = \max_{D \in T_i} \text{diam } D$ is the mesh size of T_i and C is independent of $i \in I$. Since the solution w of (23) also belongs to $H^{2+\alpha}(\Omega)$, the abstract estimate (24) combined with an error estimate for w in the H^2 -norm analogous to (66) yields the following L_2 estimate:

$$\|u - u_T\|_{L_2(\Omega)} \leq Ch_i^{2\alpha} \|u\|_{H^{2+\alpha}(\Omega)} \quad (67)$$

where C is also independent of $i \in I$.

Remark 16. The analysis of general triangular and quadrilateral C^1 macro elements can be found in Douglas *et al.* (1979).

The plate-bending problem can also be discretized by the Argyris element (cf. Example 5). If $\alpha(D) > 1$ for all $D \in \mathcal{D}$, then the nodal interpolation operator Π_D^N is well-defined for the Argyris finite element space. If $\alpha(D) \leq 1$ for some $D \in T$, then the nodal interpolation operator Π_D^N must be replaced by an interpolation operator constructed by the technique of local averaging. In either case, the estimates (65)–(67) remain valid for the Argyris finite element solution.

5 A POSTERIORI ERROR ESTIMATES AND ANALYSIS

In this section, we review explicit and implicit estimators as well as averaging and multilevel estimators for a posteriori finite element error control.

Throughout this section, we adopt the notation of Sections 2.1 and 2.2 and recall that u denotes the (unknown) exact solution of (1) while $\tilde{u} \in \tilde{V}$ denotes the discrete and given solution of (19). It is the aim of Section 5.1–5.6 to estimate the error $e := u - \tilde{u} \in V$ in the energy norm $\|\cdot\|_e = (a(\cdot, \cdot))^{1/2}$ in terms of computable quantities while Section 5.7 concerns other error norms or goal functionals.

Throughout this section, we assume $0 < \|e\|_a$ to exclude the exceptional situation $u = \tilde{u}$.

5.1 Aims and concepts in a posteriori finite element error control

The following five sections introduce the notation, the concepts of efficiency and reliability, the definitions of residual and error, a posteriori error control and adaptive algorithms, and comment on some relevant literature.

5.1.1 Error estimators, efficiency, reliability, asymptotic exactness

Regarded as an approximation to the (unknown) error norm $\|e\|_a$, a (computable) quantity η is called a *posteriori error estimator*, or *estimator* for brevity, if it is a function of the known domain Ω and its boundary Γ , the quantities of the right hand side F , cf. (6) and (13), as well as of the (given) discrete solution \tilde{u} , or the underlying triangulation.

An estimator η is called *reliable* if

$$\|e\|_a \leq C_{\text{rel}} \eta + \text{h.o.t.}_{\text{rel}} \quad (68)$$

An estimator η is called *efficient* if

$$\eta \leq C_{\text{eff}} \|e\|_a + \text{h.o.t.}_{\text{eff}} \quad (69)$$

An estimator is called *asymptotically exact* if it is reliable and efficient in the sense of (68)–(69) with $C_{\text{rel}} = C_{\text{eff}}^{-1}$.

Here, C_{rel} and C_{eff} are multiplicative constants that do not depend on the mesh size of an underlying finite element mesh T for the computation of \tilde{u} and h.o.t. denotes higher-order terms. The latter are generically much smaller than η or $\|e\|_a$, but usually, this depends on the (unknown) smoothness of the exact solution or the (known) smoothness of the given data. The readers are warned that, in general, h.o.t. may not be neglected; in case of high oscillations they may even dominate (68) or (69).

5.1.2 Error and residual

Abstract examples for estimators are (26) and (27), which involve dual norms of the residual (25). Notice carefully that $R := F - a(\tilde{u}, \cdot)$ is a bounded linear functional in V , written $R \in V^*$, and hence the dual norm

$$\|R\|_{V^*} := \sup_{v \in V \setminus \{0\}} \frac{R(v)}{\|v\|_a} = \sup_{v \in V \setminus \{0\}} \frac{a(e, v)}{\|v\|_a} = \|e\|_a < \infty \quad (70)$$

The second equality immediately follows from (25). A Cauchy inequality in (70) with respect to the scalar product

a results in $\|R\|_{V^*} \leq \|e\|_a$ while $v = e$ in (70) yields finally the equality $\|R\|_{V^*} = \|e\|_a$.

That is, the error (estimation) in the energy norm is *equivalent* to the (computation of the) dual norm of the given residual. Furthermore, it is even of *comparable computational effort* to compute an optimal $v = e$ in (70) or to compute e . The proof of (70) yields even a stability estimate: The relative error of $R(v)$ as an approximation to $\|e\|_a$ equals

$$\frac{(\|e\|_a - R(v))}{\|e\|_a} = \frac{1}{2} \left\| v - \frac{e}{\|e\|_a} \right\|_a^2 \quad \text{for all } v \in V \text{ with } \|v\|_a = 1 \quad (71)$$

In fact, given any $v \in V$ with $\|v\|_a = 1$, the identity (71) follows from

$$1 - a\left(\frac{e}{\|e\|_a}, v\right) = \frac{1}{2} a\left(\frac{e}{\|e\|_a}, \frac{e}{\|e\|_a}\right) - a\left(\frac{e}{\|e\|_a}, v\right) + \frac{1}{2} a(v, v) = \frac{1}{2} \left\| v - \frac{e}{\|e\|_a} \right\|_a^2$$

The error estimate (71) implies that the maximizing v in (70) (i.e. $v \in V$ with maximal $R(v)$ subject to $\|v\|_a \leq 1$) is unique and equals $e/\|e\|_a$. As a consequence, the computation of the maximizing v in (70) is equivalent to and indeed equally expensive as the computation of the unknown $e/\|e\|_a$ and so (since \tilde{u} is known) of the exact solution u . Therefore, a posteriori error analysis aims to compute lower and upper bounds of $\|R\|_{V^*}$ rather than its exact value.

5.1.3 Error estimators and error control

For an idealized termination procedure, one is given a tolerance $\text{Tol} > 0$ and interested in a stopping criterion (of successively adapted mesh refinements)

$$\|e\|_a \leq \text{Tol}$$

Since the error $\|e\|_a$ is unknown, it is replaced by its upper bound (68) and then leads to

$$C_{\text{rel}} \eta + \text{h.o.t.}_{\text{rel}} \leq \text{Tol} \quad (72)$$

For a verification of (72), in practice, one requires not only η but also C_{rel} and $\text{h.o.t.}_{\text{rel}}$. The latter quantity cannot be dropped; it is not sufficient to know that $\text{h.o.t.}_{\text{rel}}$ is (possibly) negligible for sufficient small mesh-sizes.

Section 5.6 presents numerical examples and further discussions of this aspect.

5.1.4 Adaptive mesh-refining algorithms

Error estimators are used in adaptive mesh-refining algorithms to motivate a *refinement rule*, which determines whether an element or edge and so on shall be refined or coarsened. This will be discussed in Section 6 below.

At this stage two remarks are in order. First, one should be precise in the language and distinguish between error estimators, which are usually global and fully involve constants and higher-order terms and (local) refinement indicators used in refinement rules. Second, constants and higher-order terms might be seen as less important and are often omitted in the usage as refinement indicators for the step MARKING in Section 6.2.

5.1.5 Literature

Amongst the most influential pioneering publications on a posteriori error control are Babuška and Rheinboldt (1978), Ladeveze and Leguillon (1983), Bank and Weiser (1985), Babuška and Miller (1987), Eriksson and Johnson (1991), followed by many others. The readers may find it rewarding to study the survey articles of Eriksson *et al.* (1995), Becker and Rannacher (2001) and the books of Verfürth (1996), Ainsworth and Oden (2000), Babuška and Strouboulis (2001), Bangert and Rannacher (2003) for a first insight and further references.

5.2 Explicit residual-based error estimators

The most frequently considered and possibly easiest class of error estimators consists of local norms of explicitly given volume and jump residuals multiplied by mesh-dependent weights.

To derive them for a general class of abstract problems from Section 2.1, let $u \in V$ be an exact solution of the problem (1) and let $\tilde{u} \in \tilde{V}$ be its Galerkin approximation from (19) with residual $R(v)$ from (25). Moreover, as in Example 1 or 2, it is supposed throughout this chapter that the strong form of the equilibration associated with the weak form (19) is of the form

$$-\text{div } p = f \quad \text{for some flux or stress } p \in L^2(\Omega; \mathbb{R}^{m \times n})$$

The discrete analog \tilde{p} is piecewise smooth but, in general, discontinuous; at several places below, it is a T piecewise constant $m \times n$ matrix as it is proportional to the gradient of some (piecewise) P_1 FE function \tilde{u} . The description of the residuals is based on the weak form of $f + \text{div } p = 0$.

5.2.1 Residual representation formula

It is the aim of this section to recast the residual in the form

$$R(v) = \sum_{T \in \mathcal{T}} \int_T r_T \cdot v \, dx - \sum_{E \in \mathcal{E}} \int_E r_E \cdot v \, ds \quad (73)$$

of a sum of integrals over all element domains $T \in \mathcal{T}$ plus a sum of integrals over all edges or faces $E \in \mathcal{E}$ and to identify the explicit volume residual r_T and the jump residual r_E .

The boundary ∂T of each finite element domain $T \in \mathcal{T}$ is a union of edges or faces, which form the set $\mathcal{E}(T)$, written $\partial T = \cup \mathcal{E}(T)$. Each edge or face $E \in \mathcal{E}$ in the set of all possible edges or faces $\mathcal{E} = \cup \{\mathcal{E}(T) : T \in \mathcal{T}\}$ is associated with a unit normal vector v_E , which is unique up to an orientation $\pm v_E$, which is globally fixed. By convention, the unit normal v on the domain Ω or on an element T points outwards.

For the ease of exploration, suppose that the underlying boundary value problem allows the bilinear form $a(\tilde{u}, v)$ to equal the sum over all $\int_T \tilde{p}_{jk} D_j v_k \, dx$ with given fluxes or stresses \tilde{p}_{jk} . Moreover, Neumann data are excluded from the description in this section and hence only interior edges contribute with a jump residual. An integration by parts on T with outer unit normal v_T yields

$$\int_T \tilde{p}_{jk} D_j v_k \, dx = \int_{\partial T} \tilde{p}_{jk} v_k v_{T,j} \, ds - \int_T v_k D_j \tilde{p}_{jk} \, dx$$

which, with the divergence operator div and proper evaluation of $\tilde{p} \cdot v$, reads

$$a(\tilde{u}, v) + \sum_{T \in \mathcal{T}} \int_T v \cdot \text{div } \tilde{p} \, dx = \sum_{T \in \mathcal{T}} \int_{\partial T} (\tilde{p} \cdot v) \cdot v \, ds$$

Each boundary ∂T is rewritten as a sum of edges or faces. Each such edge or face E belongs either to the boundary $\partial \Omega$, written $E \in \mathcal{E}_{\partial \Omega}$, or is an interior edge, written $E \in \mathcal{E}_{\Omega}$. For $E \in \mathcal{E}_{\Omega}$ there exists exactly one element T with $E \in \mathcal{E}(T)$ and one defines $T_+ = T$, $T_- = E \subset \partial \Omega$, $\omega_E = \text{int}(T)$ and $v_E := v_T = v_{\Omega}$. Any $E \in \mathcal{E}_{\Omega}$ is the intersection of exactly two elements, which we name T_+ and T_- and which essentially determine the patch $\omega_E := \text{int}(T_+ \cup T_-)$ of E . This description of T_{\pm} is unique up to the order that is fixed in the sequel by the convention that $v_E = v_{T_+}$ is exterior to T_+ . Then,

$$\sum_{T \in \mathcal{T}} \int_{\partial T} (\tilde{p} \cdot v) \cdot v \, ds = \sum_{E \in \mathcal{E}_{\Omega}} \int_E [\tilde{p} \cdot v_E] \cdot v \, ds$$

where $[\tilde{p} \cdot v_E] := (\tilde{p}|_{T_+} - \tilde{p}|_{T_-}) \cdot v_E$ for $E = \partial T_+ \cap \partial T_- \in \mathcal{E}_{\Omega}$ and $[\tilde{p} \cdot v_E] := 0$ for $E \in \mathcal{E}(T) \cap \mathcal{E}_{\partial \Omega}$. Altogether, one obtains

the error residual error representation formula (73) with the

$$\begin{aligned} \text{volume residuals } r_T &:= f + \operatorname{div} \bar{p} \quad \text{in } T \in \mathcal{T} \\ \text{jump residuals } r_E &:= [\bar{p} v_E] \quad \text{along } E \in \mathcal{E}_\Omega \end{aligned}$$

5.2.2 Weak approximation operators

In terms of the residual R , the orthogonality condition (20) is rewritten as $R(\bar{v}) = 0$ for all $\bar{v} \in \bar{V}$. Hence, given any $v \in V$ with norm $\|v\|_0 = 1$, there holds $R(v) = R(v - \bar{v})$.

Explicit error estimators rely on the design of $\bar{v} := \Pi_T^d(v)$ as a function of v . Π_T^d is called *approximation operator* as in (61)–(63) and discussed further in Section 4. See also (Carstensen, 1999; Carstensen and Funken, 2000; Nochetto and Wahlbin, 2002). For the understanding of this section, it suffices to know that there are several choices of $\bar{v} \in \bar{V}$ that satisfy first-order approximation and stability properties in the sense of

$$\begin{aligned} \sum_{T \in \mathcal{T}} \|h_T^{-1}(v - \bar{v})\|_{L_2(T)}^2 + \sum_{E \in \mathcal{E}_\Omega} \|h_E^{-1/2}(v - \bar{v})\|_{L_2(E)}^2 \\ + \|v - \bar{v}\|_{H^1(\Omega)}^2 \leq C \|v\|_{H^1(\Omega)}^2 \quad (74) \end{aligned}$$

Here, h_T and h_E denotes the diameter of an element $T \in \mathcal{T}$ and an edge $E \in \mathcal{E}$, respectively. The multiplicative constant C is independent of the mesh-sizes h_T or h_E , but depends on the shape of the element domains through their minimal angle condition (for simplices) or aspect ratio (for tensor product elements).

5.2.3 Reliability

Given the explicit volume and jump residuals r_T and r_E in (73), one defines the *explicit residual-based estimator* $\eta_{R,R}$,

$$\eta_{R,R}^2 := \sum_{T \in \mathcal{T}} h_T^2 \|r_T\|_{L_2(T)}^2 + \sum_{E \in \mathcal{E}_\Omega} h_E \|r_E\|_{L_2(E)}^2 \quad (75)$$

which is reliable, that is

$$\|e\|_a \leq C \eta_{R,R} \quad (76)$$

The proof of (76) follows from (73)–(75) and Cauchy inequalities:

$$\begin{aligned} R(v) &= R(v - \bar{v}) = \sum_{T \in \mathcal{T}} \int_T r_T \cdot (v - \bar{v}) \, dx \\ &\quad - \sum_{E \in \mathcal{E}_\Omega} \int_E r_E \cdot (v - \bar{v}) \, ds \\ &\leq \sum_{T \in \mathcal{T}} (h_T \|r_T\|_{L_2(T)} \|h_T^{-1} \|v - \bar{v}\|_{L_2(T)}) \\ &\quad + \sum_{E \in \mathcal{E}_\Omega} (h_E^{1/2} \|r_E\|_{L_2(E)} \|h_E^{-1/2} \|v - \bar{v}\|_{L_2(E)}) \end{aligned}$$

$$\begin{aligned} &+ \sum_{E \in \mathcal{E}_\Omega} (h_E^{1/2} \|r_E\|_{L_2(E)} \|h_E^{-1/2} \|v - \bar{v}\|_{L_2(E)}) \\ &\leq \left(\sum_{T \in \mathcal{T}} h_T^2 \|r_T\|_{L_2(T)}^2 \right)^{1/2} \left(\sum_{T \in \mathcal{T}} h_T^{-2} \|v - \bar{v}\|_{L_2(T)}^2 \right)^{1/2} \\ &\quad + \left(\sum_{E \in \mathcal{E}_\Omega} h_E \|r_E\|_{L_2(E)}^2 \right)^{1/2} \left(\sum_{E \in \mathcal{E}_\Omega} h_E^{-1} \|v - \bar{v}\|_{L_2(E)}^2 \right)^{1/2} \\ &\leq C \eta_{R,R} \|v\|_{H^1(\Omega)} \end{aligned}$$

For first-order finite element methods in the situation of Example 1 or 2, the volume term $r_T = f$ can be substituted by the higher-order term of oscillations, that is

$$\|e\|_a^2 \leq C \left(\operatorname{osc}(f)^2 + \sum_{E \in \mathcal{E}_\Omega} h_E \|r_E\|_{L_2(E)}^2 \right) \quad (77)$$

For each node $z \in \mathcal{N}$ with nodal basis function φ_z and patch $\omega_z := \{x \in \Omega : \varphi_z(x) \neq 0\}$ of diameter h_z and the source term $f \in L_2(\Omega)^d$ with integral mean $f_z := |\omega_z|^{-1} \int_{\omega_z} f(x) \, dx \in \mathbb{R}^d$, the oscillations of f are defined by

$$\operatorname{osc}(f) := \left(\sum_{z \in \mathcal{N}} h_z^2 \|f - f_z\|_{L_2(\omega_z)}^2 \right)^{1/2}$$

Notice for $f \in H^1(\Omega)^d$ and the mesh size $h_T \in P_0(T)$ there holds

$$\operatorname{osc}(f) \leq C \|h_T^2 Df\|_{L_2(\Omega)}$$

and so $\operatorname{osc}(f)$ is of quadratic and hence of higher-order. We refer to Carstensen and Verfürth (1999), Nochetto (1993), Becker and Rannacher (1996), and Rodriguez (1994b) for further details on and proofs of (77).

5.2.4 Efficiency

Following a technique with inverse estimates due to Verfürth (1996), this section investigates the proof of efficiency of $\eta_{R,R}$ in a local form, namely,

$$h_T \|r_T\|_{L_2(T)} \leq C (\|e\|_{H^1(T)} + \operatorname{osc}(f, T)) \quad (78)$$

$$h_E^{1/2} \|r_E\|_{L_2(E)} \leq C (\|e\|_{H^1(\omega_E)} + \operatorname{osc}(f, \omega_E)) \quad (79)$$

where \tilde{f} denotes an elementwise polynomial (best-) approximation of f and

$$\operatorname{osc}(f, T) := h_T \|f - \tilde{f}\|_{L_2(T)}$$

and

$$\operatorname{osc}(f, \omega_E) := h_E \|f - \tilde{f}\|_{L_2(\omega_E)}$$

The main tools in the proof of (79) and (78) are bubble functions b_E and b_T based on an edge or face $E \in \mathcal{E}$ and an element $T \in \mathcal{T}$ with nodes $\mathcal{N}(E)$ and $\mathcal{N}(T)$, respectively. Given a nodal basis $(\varphi_z : z \in \mathcal{N})$ of a first-order finite element method with respect to T define, for any $T \in \mathcal{T}$ and $E \in \mathcal{E}_\Omega$, the element- and edge-bubble functions

$$b_T := \prod_{z \in \mathcal{N}(T)} \varphi_z \in H_0^1(T) \quad \text{and} \quad b_E := \prod_{z \in \mathcal{N}(E)} \varphi_z \in H_0^1(\omega_E) \quad (80)$$

b_E and b_T are nonnegative and continuous piecewise polynomials ≤ 1 with support $\operatorname{supp} b_E = \bar{\omega}_E = T_+ \cup T_-$ (for $T_\pm \in \mathcal{T}$ with $E = T_+ \cap T_-$) and $\operatorname{supp} b_T = T$.

Utilizing the bubble functions (80), the proof of (78)–(79) essentially consists in the design of test functions $w_T \in H_0^1(T)$, $T \in \mathcal{T}$, and $w_E \in H_0^1(\omega_E)$, $E \in \mathcal{E}_\Omega$, with the properties

$$|w_T|_{H^1(T)} \leq C h_T \|r_T\|_{L_2(T)}$$

and

$$|w_E|_{H^1(\omega_E)} \leq C h_E^{1/2} \|r_E\|_{L_2(E)} \quad (81)$$

$$h_T^2 \|r_T\|_{L_2(T)}^2 \leq C_1 R(w_T) + C_2 \operatorname{osc}(f, T)^2 \quad (82)$$

$$h_E \|r_E\|_{L_2(E)}^2 \leq C_1 R(w_E) + C_2 \operatorname{osc}(f, \omega_E)^2 \quad (83)$$

In fact, (81)–(83), the definition of the residual $R = a(e, \cdot)$, and Cauchy inequalities with respect to the scalar product a prove (78)–(79).

To construct the test function w_T , $T \in \mathcal{T}$, recall $\operatorname{div} p + f = 0$ and $r_T = f + \operatorname{div} \bar{p}$ and set $\tilde{r}_T := \tilde{f} + \operatorname{div} \tilde{p}$ for some polynomial \tilde{f} on T such that \tilde{r}_T is a best approximation of r_T in some finite-dimensional (polynomial) space with respect to $L_2(T)$. Since

$$\begin{aligned} h_T \| \tilde{r}_T \|_{L_2(T)} &\leq h_T \| r_T \|_{L_2(T)} \leq h_T \| \tilde{r}_T \|_{L_2(T)} \\ &\quad + h_T \| f - \tilde{f} \|_{L_2(T)} \end{aligned}$$

it remains to bound \tilde{r}_T , which belongs to a finite-dimensional space and hence satisfies an inverse inequality

$$h_T \| \tilde{r}_T \|_{L_2(T)} \leq C h_T \| b_T^{1/2} \tilde{r}_T \|_{L_2(T)}$$

This motivates the estimation of

$$\begin{aligned} \| b_T^{1/2} \tilde{r}_T \|_{L_2(T)}^2 &= \int_T b_T \tilde{r}_T \cdot (\tilde{r}_T - r_T) \, dx + \int_T b_T \tilde{r}_T \cdot r_T \, dx \\ &\leq \| b_T^{1/2} \tilde{r}_T \|_{L_2(T)} \| b_T^{1/2} (f - \tilde{f}) \|_{L_2(T)} \\ &\quad + \int_T b_T \tilde{r}_T \cdot \operatorname{div} (\tilde{p} - p) \, dx \end{aligned}$$

The combination of the preceding estimates results in

$$\begin{aligned} h_T^2 \| r_T \|_{L_2(T)}^2 &\leq C_1 \int_T (h_T^2 b_T \tilde{r}_T \cdot \operatorname{div} (\tilde{p} - p) \, dx \\ &\quad + C_2 \operatorname{osc}(f, T)^2 \end{aligned}$$

An integration by parts concludes the proof of (82) for

$$w_T := h_T^2 b_T \tilde{r}_T \quad (84)$$

the proof of (81) for this w_T is immediate.

Given an interior edge $E = T_+ \cap T_- \in \mathcal{E}_\Omega$ with its neighboring elements T_+ and T_- , simultaneously addressed as $T_\pm \in \mathcal{T}$, extend the edge residual r_E from the edge E to its patch $\omega_E = \operatorname{int}(T_+ \cup T_-)$ such that

$$\begin{aligned} \| b_E r_E \|_{L_2(\omega_E)} + h_E \| b_E r_E \|_{H^1(\omega_E)} &\leq C_1 h_E^{1/2} \| r_E \|_{L_2(E)} \\ &\leq C_2 h_E^{1/2} \| b_E^{1/2} r_E \|_{L_2(E)} \end{aligned} \quad (85)$$

(with an inverse inequality at the end). The choice of the two real constants

$$\alpha_\pm = \frac{\int_{T_\pm} h_E b_E \tilde{r}_{T_\pm} \cdot r_E \, dx}{\int_{T_\pm} w_{T_\pm} \cdot \tilde{r}_{T_\pm} \, dx}$$

in the definition

$$w_E := \alpha_+ w_{T_+} + \alpha_- w_{T_-} - h_E b_E r_E \quad (86)$$

yields $\int_{T_\pm} w_{T_\pm} \cdot \tilde{r}_{T_\pm} \, dx = 0$. Since $\int_{T_\pm} w_{T_\pm} \cdot \tilde{r}_{T_\pm} \, dx = h_{T_\pm}^2 \| b_{T_\pm}^{1/2} \tilde{r}_{T_\pm} \|_{L_2(T_\pm)}^2$, one eventually deduces $|\alpha_\pm| |w_{T_\pm}|_{H^1(T_\pm)} \leq C h_E^{1/2} \| r_E \|_{L_2(E)}$ and then concludes (81). An integration by parts shows

$$\begin{aligned} C^{-2} h_E \| r_E \|_{L_2(E)}^2 &\leq h_E \| b_E^{1/2} r_E \|_{L_2(E)}^2 \\ &= - \int_E w_E \cdot r_E \, ds = \int_E w_E \cdot [(p - \bar{p}) \cdot v_E] \, ds \\ &= \int_{\omega_E} (p - \bar{p}) : D w_E \, dx + \int_{\omega_E} w_E \cdot \operatorname{div}_T (p - \bar{p}) \, dx \\ &= R(w_E) - \int_{\omega_E} w_E \cdot (f + \operatorname{div}_T \bar{p}) \, dx \\ &= R(w_E) - \int_{\omega_E} w_E \cdot (f - \tilde{f}) \, dx \end{aligned}$$

(with $\int_{T_\pm} w_{T_\pm} \cdot \tilde{r}_{T_\pm} \, dx = 0$ in the last step). A Friedrichs inequality $\|w_E\|_{L_2(\omega_E)} \leq C h_E |w_E|_{H^1(\omega_E)}$ and (81) then conclude the proof of (83).

5.3 Implicit error estimators

Implicit error estimators are based on a local norm of a solution of a localized problem of a similar type with the residual terms on the right-hand side. This section

introduces two different versions based on a partition of unity and based on an equilibration technique.

5.3.1 Localization by partition of unity

Given a nodal basis $(\varphi_z : z \in \mathcal{N})$ of a first-order finite element method with respect to T , there holds the partition of unity property

$$\sum_{z \in \mathcal{N}} \varphi_z = 1 \quad \text{in } \Omega$$

Given the residual $R = F - a(\tilde{u}, \cdot) \in V^*$, we observe that $R_z(v) := R(\varphi_z v)$ defines a bounded linear functional R_z on a localized space called V_z and specified below.

The bilinear form a is an integral over ω on some integrand. The latter may be weighted with φ_z to define some (localized) bilinear form $a_z : V_z \times V_z \rightarrow \mathbb{R}$. Supposing that a_z is V_z -elliptic one defines the norm $\|\cdot\|_{a_z}$ on V_z and considers

$$\eta_z := \sup_{v \in V_z \setminus \{0\}} \frac{R_z(v)}{\|v\|_{a_z}} \quad (87)$$

The dual norm is as in (70)–(71) and hence equivalent to the computation of the norm $\|e_z\|_{a_z}$ of a local solution

$$e_z \in V_z \quad \text{with} \quad a_z(e_z, \cdot) = R_z \in V_z^* \quad (88)$$

(The proof of $\|e_z\|_{a_z} = \eta_z$ follows the arguments of Section 5.1.2 and hence is omitted.)

Example 10. Adopt notation from Example 1 and let $(\varphi_z : z \in \mathcal{N})$ be the first-order finite element nodal basis functions. Then define R_z and a_z by

$$R_z(v) := \int_{\Omega} \varphi_z f v \, dx - \int_{\Omega} \nabla \tilde{u} \cdot \nabla(\varphi_z v) \, dx \quad \forall v \in V$$

$$a_z(v_1, v_2) := \int_{\Omega} \varphi_z \nabla v_1 \cdot \nabla v_2 \, dx \quad \forall v_1, v_2 \in V$$

Let V_z denote the completion of V under the norm given by the scalar product a_z when $\varphi_z \neq 0$ on Γ or otherwise its quotient space with \mathbb{R} , i.e.

$$V_z = \begin{cases} \{v \in H_{\text{loc}}^1(\omega_z) : a_z(v, v) < \infty, \varphi_z v = 0 \text{ on } \Gamma \cap \partial\omega_z\} & \text{if } \varphi_z \neq 0 \text{ on } \Gamma \\ \{v \in H_{\text{loc}}^1(\omega_z) : a_z(v, v) < \infty, \int_{\Omega} \varphi_z v \, dx = 0\} & \text{if } \varphi_z \equiv 0 \text{ on } \Gamma \end{cases}$$

Notice that $R_z(1) = 0$ for a free node z such that (88) has a unique solution and hence $\eta_z < \infty$.

In practical applications, the solution e_z of (88) has to be approximated by some finite element approximation \tilde{e}_z

on a discrete space \tilde{V}_z based on a finer mesh or of higher order. (Arguing as in the stability estimate (71), leads to an error estimate for an approximation $\|\tilde{e}_z\|_{a_z}$ of η_z .)

Suppose that η_z is known exactly (or computed with high and controlled accuracy) and that the bilinear form a is localized through the partition of unity such that (e.g. in Example 10)

$$a(u, v) = \sum_{z \in \mathcal{N}} a_z(u, v) \quad \forall u, v \in V \quad (89)$$

Then the implicit error estimator η_L is reliable with $C_{\text{rel}} = 1$ and $\text{h.o.t.}_{\text{rel}} = 0$,

$$\|e\|_a \leq \eta_L := \left(\sum_{z \in \mathcal{N}} \eta_z^2 \right)^{1/2} \quad (90)$$

The proof of (90) follows from the definition of R_z , η_z , and e_z and Cauchy inequalities:

$$\begin{aligned} \|e\|_a^2 &= R(e) = \sum_{z \in \mathcal{N}} R_z(e) \leq \sum_{z \in \mathcal{N}} \eta_z \|e\|_{a_z} \\ &\leq \left(\sum_{z \in \mathcal{N}} \eta_z^2 \right)^{1/2} \left(\sum_{z \in \mathcal{N}} \|e\|_{a_z}^2 \right)^{1/2} = \eta_L \|e\|_a \end{aligned}$$

Notice that $\|\tilde{e}_z\|_{a_z} := \tilde{\eta}_z \leq \eta_z$ for any approximated local solution

$$\tilde{e}_z \in \tilde{V}_z \quad \text{with} \quad a_z(\tilde{e}_z, \cdot) = R_z \in \tilde{V}_z^* \quad (91)$$

and all of them are efficient estimators. The proof of efficiency is based on a weighted Poincaré or Friedrichs inequality which reads

$$\|\varphi_z v\|_a \leq C \|v\|_{a_z} \quad \forall v \in V_z \quad (92)$$

In fact, in Example 1, 2, and 3, one obtains efficiency in a more local form than indicated in

$$\eta_z \leq C \|e\|_a \quad \text{with h.o.t.}_{\text{eff}} = 0 \quad (93)$$

(This follows immediately from (92):

$$\begin{aligned} R_z(v) &= R(\varphi_z v) = a(e, \varphi_z v) \leq \|e\|_a \|\varphi_z v\|_a \\ &\leq C \|e\|_a \|v\|_{a_z} \end{aligned}$$

In the situation of Example 10, the estimator η_L dates back to Babuška and Miller (1967); the use of weights was established in Carstensen and Funken (1999/00). A reliable computable estimator $\tilde{\eta}_L$ is introduced in Morin, Nochetto

and Siebert (2003a) based on a proper finite-dimensional space \tilde{V}_z of some piecewise quadratic polynomials on ω_z .

5.3.2 Equilibration estimators

The nonoverlapping domain decomposition schemes employ artificial unknowns $g_T \in L_2(\partial T)^m$ for each $T \in \mathcal{T}$ at the interfaces, which allow a representation of the form

$$R(v) = \sum_{T \in \mathcal{T}} R_T(v) \quad \text{where}$$

$$R_T(v) := \int_T f \cdot v \, dx - \int_T \tilde{p} : Dv \, dx + \int_{\partial T} g_T \cdot v \, ds \quad (94)$$

Adopting the notation from Section 5.2.1, the new quantities g_T satisfy

$$g_{T_+} + g_{T_-} = 0 \quad \text{along } E = \partial T_+ \cap \partial T_- \in \mathcal{E}_{\text{int}}$$

(where T_+ and T_- denote neighboring element domains) to guarantee (94). (There are non-displayed modifications on any Neumann boundary edge $E \subset \partial\Omega$.) Moreover, the bilinear form a is expanded in an elementwise form

$$a(u, v) = \sum_{T \in \mathcal{T}} a_T(u, v) \quad \forall u, v \in V \quad (95)$$

Under the *equilibration condition* $R_T(c) = 0$ for all kernel functions c (namely, the constant functions for the Laplace model problem), the resulting *local problem* reads

$$e_T \in V_T \quad \text{with } a_T(e_T, \cdot) = R_T \in V_T^* \quad (96)$$

and is equivalent to the computation of

$$\eta_T := \sup_{v \in V_T \setminus \{0\}} \frac{R_T(v)}{\|v\|_{a_T}} = \|e_T\|_{a_T} \quad (97)$$

The sum of all local contributions defines the reliable *equilibration error estimator* $\eta_{\mathcal{G}}$,

$$\|e\|_a \leq \eta_{\mathcal{G}} := \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{1/2} \quad (98)$$

(The immediate proof of (98) is analogous to that of (90) and hence is omitted.)

Example 11. In the situation of Example 10 there holds $\eta_T < \infty$ if and only if either $\Gamma \cap \partial T$ has positive surface measure (with $V_T = \{v \in H^1(T) : v = 0 \text{ on } \Gamma \cap \partial T\}$) or otherwise $R_T(1) = 0$ (with $V_T = \{v \in H^1(T) : \int_T v \, dx = 0\}$). Ladeveze and Leguillon (1983) suggested a certain choice of the interface corrections to guarantee this and

even higher-order equilibrations are established. Details on the implementation are given in Ainsworth and Oden (2000); a detailed error analysis with higher-order equilibrations and the perturbation by a finite element simulation of the local problems with corrections can be found in Ainsworth and Oden (2000) and Babuška and Strouboulis (2001).

The error estimator $\eta = \eta_{\mathcal{G}}$ is efficient in the sense of (69) with higher-order terms $\text{h.o.t.}(T)$ on T that depend on the given data provided

$$h_E^{1/2} \|g_T - \tilde{p} v_E\|_{L_2(E)} \leq C \|e\|_{a_T} + \text{h.o.t.}(T) \quad \text{for all } E \in \mathcal{E}(T) \quad (99)$$

(Recall that $\mathcal{E}(T)$ denotes the set of edges or faces of T .) This stability property depends on the design of g_T ; a positive example is given in Theorem 6.2 of Ainsworth and Oden (2000) for Example 1. Given Inequality (99), the efficiency of η_T follows with standard arguments, for example, an integration by parts, a trace and Poincaré or Friedrichs inequality $h_T^{-1} \|v\|_{L_2(T)} + h_T^{-1/2} \|v\|_{L_2(\partial T)} \leq C \|v\|_{a_T}$ for $v \in V_T$:

$$\begin{aligned} R_T(v) &= \int_T r_T \cdot v \, dx + \int_{\partial T} (g_T - \tilde{p} v) \cdot v \, ds \\ &\leq C \left(h_T \|r_T\|_{L_2(T)} + h_T^{1/2} \|g_T - \tilde{p} v\|_{L_2(\partial T)} \right) \|v\|_{a_T} \end{aligned}$$

followed by (79) and (99).

5.4 Multilevel error estimators

While the preceding estimators evaluate or estimate the residual of one finite element solution u_H , multilevel estimators concern at least two meshes \mathcal{T}_H and \mathcal{T}_h with associated discrete spaces $V_H \subset V_h \subset V$ and two discrete solutions $u_H = \tilde{u}$ and u_h . The interpretation is that p_h is computed on a finer mesh (e.g. \mathcal{T}_h is a refinement of \mathcal{T}_H) or that p_h is computed with higher polynomial order than $p_H = \tilde{p}$.

5.4.1 Error-reduction property and multilevel error estimator

Let $V_H \subset V_h \subset V$ denote two nested finite element spaces in V with coarse and fine finite element solution $u_H = \tilde{u} \in V_H = \tilde{V}$ and $u_h \in V_h$ of the discrete problem (19), respectively, and with the exact solution u . Let $p_H = \tilde{p}$, p_h , and p denote the respective fluxes and let $\|\cdot\|$ be a norm associated to the energy norm, for example,

a norm with $\|p - \tilde{p}\| = \|u - \tilde{u}\|_a$ and $\|p - p_h\| = \|u - u_h\|_a$. Then, the *multilevel error estimator*

$$\eta_{ML} := \|p_h - p_H\| = \|u_h - u_H\|_a \quad (100)$$

simply is the norm of the difference of the two discrete solutions. The interpretation is that the error $\|p - p_h\|$ of the finer discrete solution is systematically smaller than the error $\|e\|_a = \|p - p_H\|$ of the coarser discrete solution in the sense of an *error-reduction property*: For some constant $\varrho < 1$, there holds

$$\|p - p_h\| \leq \varrho \|p - p_H\| \quad (101)$$

Notice the bound $\varrho \leq 1$ for Galerkin errors in the energy norm (because of the best-approximation property). The point is that $\varrho < 1$ in (101) is bounded away from one. Then, the error-reduction property (101) immediately implies reliability and efficiency of η_{ML} :

$$(1 - \varrho) \|p - p_H\| \leq \eta_{ML} = \|p_h - p_H\| \leq (1 + \varrho) \|p - p_H\| \quad (102)$$

(The immediate proof of (102) utilizes (101) and a simple triangle inequality.)

Four remarks on the error-reduction conclude this section: Efficiency of η_{ML} in the sense of (69) is robust in $\varrho \rightarrow 1$, but reliability is not: The reliability constant $C_{rel} = (1 - \varrho)^{-1}$ in (68) tends to infinity as ϱ approaches 1.

Higher-order terms are invisible in (102): $\text{h.o.t.}_{rel} = 0 = \text{h.o.t.}_{eff}$. This is unexpected when compared to all the other error estimators and hence indicates that (101) should fail to hold for heavily oscillating right-hand sides.

The error-reduction property (101) is often observed in practice for fine meshes and can be monitored during the calculation. For coarse meshes and in the preasymptotic range, (101) may fail to hold.

The error-reduction property (101) is often called *saturation assumption* in the literature and frequently has a status of an unproven hypothesis.

5.4.2 Counterexample for error-reduction

The error-reduction property (101) may fail to hold even if f shows no oscillations: Figure 10 displays two triangulations, \mathcal{T}_H and its refinement \mathcal{T}_h , with one and five free nodes, respectively. If the right-hand side is constant and if the problem has homogeneous Dirichlet conditions for the Poisson problem

$$1 + \Delta u = 0 \text{ in } \Omega := (0, 1)^2 \text{ and } u = 0 \text{ on } \partial\Omega$$

then the corresponding P_1 finite element solutions coincide: $u_H = u_h$. A direct proof is based on the nodal basis function

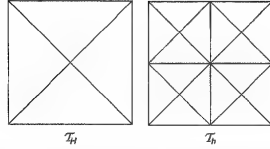


Figure 10. Two triangulations \mathcal{T}_H and \mathcal{T}_h with equal discrete P_1 finite element solutions $u_H = u_h$ for a Poisson problem with right-hand side $f = 1$. The refinement \mathcal{T}_h of \mathcal{T}_H is generated by two (newest-vertex) bisections per element (initially with the interior node in \mathcal{T}_h as newest vertex).

Φ_1 of the free node in $V_H := FE_{\mathcal{T}_H}$ (the first-order finite element space with respect to the coarse mesh \mathcal{T}_H) and the nodal basis functions $\Phi_2, \dots, \Phi_5 \in V_h := S_0^1(\mathcal{T}_h)$ of the new free nodes in \mathcal{T}_h . Then,

$$\Phi_2 := \Phi_1 - (\varphi_2 + \dots + \varphi_5) \in V_h := FE_{\mathcal{T}_h}$$

satisfies (since $\int_E \Phi_2 dx = 0$ for all edges E in \mathcal{T}_H and $\int_\Omega \Phi_2 dx = 0$)

$$R(\Phi_2) = 0$$

Thus u_H is the finite element solution in $W_h := \text{span}\{\Phi_1, \Phi_2\} \subset V_h$. Since, by symmetry, the finite element solution u_h in V_h belongs to W_h , there holds $u_H = u_h$.

5.4.3 Affirmative example for error-reduction

Adopt notation from Section 5.2.4 with a coarse discrete space $V_H = \bar{V}$ and consider the fine space $V_h := V_H \oplus W_h$ for

$$W_h := \text{span}\{\tilde{r}_T b_T : T \in \mathcal{T}\} \oplus \text{span}\{r_E b_E : E \in \mathcal{E}_\Omega\} \subset V \quad (103)$$

Then there holds the error-reduction property up to higher-order terms

$$\text{osc}(f) := \left(\sum_{T \in \mathcal{T}} h_T^2 \|f - \tilde{f}\|_{L_2(T)}^2 \right)^{1/2}$$

namely

$$\|p - p_h\|^2 \leq \varrho \|p - p_H\|^2 + \text{osc}(f)^2 \quad (104)$$

The constant ϱ in (104) is uniformly smaller than one, independent of the mesh size, and depends on the shape of the elements and the type of ansatz functions through constants in (81)–(83).

The proof of (104) is based on the test functions w_T and w_E in (84) and (86) of Section 5.2.4 and

$$w_h := \sum_{T \in \mathcal{T}} w_T + \sum_{E \in \mathcal{E}_\Omega} w_E \in W_h \subset V$$

Utilizing (81)–(83) one can prove

$$\begin{aligned} \|u_h\|_a^2 &\leq C \left(\sum_{T \in \mathcal{T}} h_T^2 \|\tilde{r}_T\|_{L_2(T)}^2 + \sum_{E \in \mathcal{E}_\Omega} h_E \|r_E\|_{L_2(E)}^2 \right) \\ &\leq C \left(\sum_{T \in \mathcal{T}} R(w_T) + \sum_{E \in \mathcal{E}_\Omega} R(w_E) + \text{osc}(f)^2 \right) \\ &= C(R(u_h) + \text{osc}(f)^2) \end{aligned}$$

Since w_h belongs to V_h and u_h is the finite element solution with respect to V_h there holds

$$R(w_h) = a(u_h - u_H, w_h) \leq \|u_h - u_H\|_a \|w_h\|_a$$

The combination of the preceding inequalities yields the key inequality

$$\begin{aligned} \eta_{h,R}^2 &= \sum_{T \in \mathcal{T}} h_T^2 \|\tilde{r}_T\|_{L_2(T)}^2 + \sum_{E \in \mathcal{E}_\Omega} h_E \|r_E\|_{L_2(E)}^2 \\ &\leq C(\|u_h - u_H\|_a^2 + \text{osc}(f)^2) \end{aligned}$$

This, the Galerkin orthogonality $\|u - u_H\|_a^2 = \|u - u_h\|_a^2 + \|u_h - u_H\|_a^2$, and the reliability of η_h show

$$\|u - u_H\|_a^2 \leq C C_{rel}^2 (\|u - u_H\|_a^2 - \|u - u_h\|_a^2 + \text{osc}(f)^2)$$

and so imply (104) with $\varrho = 1 - C^{-1} C_{rel}^{-2} < 1$.

Example 12. In the Poisson problem with P_1 finite elements and discrete space V_H , let W_h consist of the quadratic and cubic bubble functions (80). Then (104) holds; cf. also Example 14 below.

Example 13. Other affirmative examples for the Poisson problem consist of the P_1 and P_2 finite element spaces V_H and V_h over one regular triangulation \mathcal{T} or of the P_1 finite element spaces with respect to a regular triangulation \mathcal{T}_H and its red-refinement \mathcal{T}_h . The observation that the element-bubble functions are in fact redundant is due to Dörfler and Nochetto (2002).

5.4.4 Hierarchical error estimator

Given a smooth right-hand side f and based on the example of the previous section, the multilevel estimator (100)

is reliable and efficient up to higher-order terms. The costly calculation of u_h , however, exclusively allows for an accurate error control of u_H (and no reasonable error control for u_h). Instead of (100), cheaper versions are favored where u_h is replaced by some quantity computed by a localized problem. One reliable and efficient version is the *hierarchical error estimator*

$$\eta_H := \left(\sum_{T \in \mathcal{T}} \eta_T^2 + \sum_{E \in \mathcal{E}} \eta_E^2 \right)^{1/2} \quad (105)$$

where, for each $T \in \mathcal{T}$ and $E \in \mathcal{E}$ and their test functions (84) and (86),

$$\eta_T := \frac{R(w_T)}{\|w_T\|_a} \quad \text{and} \quad \eta_E := \frac{R(w_E)}{\|w_E\|_a} \quad (106)$$

(The proof of reliability and efficiency follows from (81)–(83) by the arguments from Section 5.2.4.)

Example 14. In the Poisson problem with P_1 finite elements and discrete space V_H , let W_h consist of the quadratic and cubic bubble functions (80). Then,

$$\eta_H := \left(\sum_{T \in \mathcal{T}} \frac{R(b_T)^2}{\|b_T\|_a^2} + \sum_{E \in \mathcal{E}_\Omega} \frac{R(b_E)^2}{\|b_E\|_a^2} \right)^{1/2} \quad (107)$$

is a reliable and efficient hierarchical error estimator. With the error-reduction property of P_1 and P_2 finite elements due to Dörfler and Nochetto (2002),

$$\eta_H := \left(\sum_{E \in \mathcal{E}_\Omega} \frac{R(b_E)^2}{\|b_E\|_a^2} \right)^{1/2} \quad (108)$$

is reliable and efficient as well. The same is true if, for each edge $E \in \mathcal{E}$, b_E defines a hat function of the midpoint of E with respect to a red-refinement \mathcal{T}_h of \mathcal{T}_H (that is, each edge is halved and each triangle is divided into four congruent subtriangles; cf. Figure 15, left).

5.5 Averaging error estimators

Averaging techniques, also called (*gradient*) *recovery estimators*, focus on one mesh and one known low-order flux approximation \tilde{p} and the difference to a piecewise polynomial \tilde{q} in a finite-dimensional subspace $\tilde{Q} \subset L_2(\Omega; \mathbb{R}^{m \times n})$ of higher polynomial degrees and more restrictive continuity conditions than those generally satisfied by \tilde{p} . Averaging techniques are universal in the sense that there is no need for any residual or partial differential equation in order to apply them.

5.5.1 Definition of averaging error estimators

The procedure consists of taking a piecewise smooth \tilde{p} and approximating it by some globally continuous piecewise polynomials (denoted by \bar{Q}) of higher degree $A\tilde{p}$. A simple example, frequently named after Zienkiewicz and Zhu and sometimes even called the ZZ estimator, reads as follows: For each node $z \in \mathcal{N}$ and its patch ω_z let

$$(A\tilde{p})(z) = \frac{\int_{\omega_z} \tilde{p} \, dx}{\int_{\omega_z} 1 \, dx} \in \mathbb{R}^{m \times n} \quad (109)$$

be the integral mean of \tilde{p} over ω_z . Then, define $A\tilde{p}$ by interpolation with (conforming, i.e. globally continuous) hat functions φ_z , for $z \in \mathcal{N}$,

$$A\tilde{p} = \sum_{z \in \mathcal{N}} (A\tilde{p})(z) \varphi_z \in \bar{Q}$$

Let $\bar{Q} = \text{span}\{\varphi_z; z \in \mathcal{N}\}$ denote the (conforming) first-order finite element space and let $\|\cdot\|$ be the norm for the fluxes. Then the averaging estimator is defined by

$$\eta_A := \|\tilde{p} - A\tilde{p}\| \quad (110)$$

Notice that there is a minimal version

$$\eta_M := \min_{\tilde{q} \in \bar{Q}} \|\tilde{p} - \tilde{q}\| \leq \eta_A \quad (111)$$

The efficiency of η_M follows from a triangle inequality, namely

$$\eta_M \leq \|p - \tilde{p}\| + \|p - \tilde{p}\| \quad \text{for all } \tilde{q} \in \bar{Q} \quad (112)$$

and the fact that $\|p - \tilde{p}\| = \mathcal{O}(h)$ while (in all the examples of this chapter)

$$\min_{\tilde{q} \in \bar{Q}} \|p - \tilde{q}\| = \text{h.o.t.}(p) =: \text{h.o.t.}_{\text{eff}}$$

This is of higher order for smooth p and efficiency follows for $\eta = \eta_M$ and $C_{\text{eff}} = 1$.

It turns out that η_A and η_M are very close and accurate estimators in many numerical examples; cf. Section 5.6.4 below. This and the fact that the calculation of η_A is an easy postprocessing made η_A extremely popular.

For proper treatment of Neumann boundary conditions, we refer to Carstensen and Bartels (2002) and for applications in solid and fluid mechanics to Alberty and Carstensen (2003) and Carstensen and Funken (2001a,b) and for higher-order FEM to Bartels and Carstensen (2002).

Multigrid smoothing steps may be successfully employed as averaging procedures as proposed in Bank and Xu (2003).

5.5.2 All averaging error estimators are reliable

The first proof of reliability dates back to Rodríguez (1994a,b) and we refer to Carstensen (2004) for an overview. A simplified reliability proof for η_M and hence for all averaging techniques (Carstensen, Bartels and Klose, 2001) is outlined in the sequel.

First let Π be the L_2 projection onto the first-order finite element space \bar{V} and let \tilde{q} be arbitrary in \bar{Q} , that is, each of the $m \times n$ components of \tilde{q} is a first-order finite element function in $F\bar{E}_T$. The Galerkin orthogonality shows for the error $e := u - \tilde{u}$ and $\tilde{p} := \nabla \tilde{u}$ in the situation of Example 1 that

$$\begin{aligned} \|e\|_a^2 &= \int_{\Omega} (\nabla u - \tilde{q}) \cdot \nabla (e - \Pi e) \, dx \\ &\quad + \int_{\Omega} (\tilde{q} - \tilde{p}) \cdot \nabla (e - \Pi e) \, dx \end{aligned}$$

A Cauchy inequality in the latter term is combined with the H^1 -stability of Π , namely,

$$\|\nabla(e - \Pi e)\|_{L_2(\Omega)} \leq C_{\text{stab}} \|\nabla e\|_{L_2(\Omega)}$$

(For sufficient conditions for this we refer to Crouzeix and Thomée (1987), Bramble, Pasciak and Steinbach (2002), Carstensen (2002, 2003b), and Carstensen and Verfürth (1999).) The H^1 -stability in the second term and an integration by parts in the first term on the right-hand side show

$$\begin{aligned} \|e\|_a^2 &\lesssim \int_{\Omega} f \cdot (e - \Pi e) \, dx + \int_{\Omega} (e - \Pi e) \cdot \text{div } \tilde{q} \, dx \\ &\quad + C_{\text{stab}} \|\nabla e\|_{L_2(\Omega)} \|\tilde{p} - \tilde{q}\|_{L_2(\Omega)} \end{aligned}$$

Since $e - \Pi e$ is L_2 -orthogonal onto $f_h := \Pi f \in \bar{V}$,

$$\begin{aligned} \int_{\Omega} f \cdot (e - \Pi e) \, dx &= \int_{\Omega} (f - f_h) \cdot (e - \Pi e) \, dx \\ &\leq \|h_T^{-1}(e - \Pi e)\|_{L_2(\Omega)} \|h_T(f - \Pi f)\|_{L_2(\Omega)} \end{aligned}$$

Notice that, despite possible boundary layers, $\|h_T(f - \Pi f)\|_{L_2(\Omega)}$ is h.o.t. of higher order. The first-order approximation property of the L_2 projection,

$$\|h_T^{-1}(e - \Pi e)\|_{L_2(\Omega)} \leq C_{\text{approx}} \|\nabla e\|_{L_2(\Omega)}$$

follows from the H^1 -stability (cf. e.g. Carstensen and Verfürth (1999) for a proof). Similar arguments for the remaining term $\text{div } \tilde{q}$ and the T -piecewise divergence operator div_T with $\text{div } \tilde{q} = \text{div}_T \tilde{q} = \text{div}_T(\tilde{q} - \tilde{p})$ (recall that \tilde{u} is of first-order and hence $\Delta \tilde{u} = 0$ on each element) lead to

$$\begin{aligned} \int_{\Omega} (e - \Pi e) \cdot \text{div } \tilde{q} \, dx &\leq C_{\text{approx}} \|\nabla e\|_{L_2(\Omega)} \|h_T \text{div}_T(\tilde{q} - \tilde{p})\|_{L_2(\Omega)} \end{aligned}$$

An inverse inequality $h_T \|\text{div}_T(\tilde{q} - \tilde{p})\|_{L_2(T)} \leq C_{\text{inv}} \|\tilde{q} - \tilde{p}\|_{L_2(T)}$ (cf. Section 3.4) shows

$$\int_{\Omega} (e - \Pi e) \cdot \text{div } \tilde{q} \, dx \leq C_{\text{approx}} C_{\text{inv}} \|\nabla e\|_{L_2(\Omega)} \|\tilde{q} - \tilde{p}\|_{L_2(\Omega)}$$

The combination of all established estimates plus a division by $\|e\|_a = \|\nabla e\|_{L_2(\Omega)}$ yield the announced reliability result

$$\|e\|_a \leq (C_{\text{stab}} + C_{\text{approx}} C_{\text{inv}}) \|\tilde{p} - \tilde{q}\|_{L_2(\Omega)} + \text{h.o.t.}$$

In the second step, one designs a more local approximation operator J to substitute Π as in Carstensen and Bartels (2002); the essential properties are the H^1 -stability, the first-order approximation property, and a local form of the orthogonality condition; we omit the details.

5.5.3 Averaging error estimators and edge contributions

There is a local equivalence of the estimators η_A from a local averaging process (109) and the edge estimator

$$\eta_E := \left(\sum_{E \in \mathcal{E}_h} h_E \|\tilde{p} \nu_E\|_{L_2(E)}^2 \right)^{1/2}$$

The observation that, with some mesh-size-independent constant C ,

$$C^{-1} \eta_E \leq \eta_A \leq C \eta_E \quad (113)$$

dates back to Rodríguez (1994a) and can be found in Verfürth (1996). The proof of (113) for piecewise linears in \bar{V} is based on the equivalence of the two seminorms

$$Q_1(\tilde{q}) := \min_{r \in \mathbb{R}^{m \times n}} \|\tilde{q} - r\|_{L_2(\omega_z)}$$

and

$$Q_2(\tilde{q}) := \left(\sum_{E \in \mathcal{E}_z} h_E \|\tilde{q} \nu_E\|_{L_2(E)}^2 \right)^{1/2}$$

for piecewise constant vector-valued functions \tilde{q} in P_z , the set of possible fluxes \tilde{q} restricted on the patch ω_z of a

node z , and with the set of edges $\mathcal{E}_z := \{E \in \mathcal{E} : z \in E\}$. The main observation is that Q_1 and Q_2 vanish exactly for constants functions $\mathbb{R}^{m \times n}$ and hence they are norms on the quotient space $P_z/\mathbb{R}^{m \times n}$. By the equivalence of norms on any finite-dimensional space $P_z/\mathbb{R}^{m \times n}$, there holds

$$C^{-1} Q_1(\tilde{q}) \leq Q_2(\tilde{q}) \leq C Q_1(\tilde{q}) \quad \forall \tilde{q} \in P_z$$

This is a local version of (113), which eventually implies (113) by localization and composition; we omit the details.

For triangles and tetrahedra and piecewise linear finite element functions, it is proved in Carstensen (2004) that

$$\eta_M \leq \eta_A \leq C_d \eta_M \quad (114)$$

with universal constants $C_2 = \sqrt{10}$ and $C_3 = \sqrt{15}$ for 2-D and 3-D, respectively. This equivalence holds for a larger class of elements and (first-order) averaging operators and then proves efficiency for η_A whereas efficiency of η_M follows from a triangle inequality in (112).

5.6 Comparison of error bounds in benchmark example

This section is devoted to a numerical comparison of energy errors and its a posteriori error estimators for an elliptic model problem.

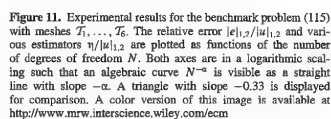
5.6.1 Benchmark example

The numerical comparisons are computed for the Poisson problem

$$1 + \Delta u = 0 \text{ in } \Omega \quad \text{and} \quad u = 0 \text{ on } \partial\Omega \quad (115)$$

on the L-shaped domain $\Omega = (-1, +1)^2 \setminus ([0, 1] \times [-1, 0])$ and its boundary $\partial\Omega$. The first mesh \mathcal{T}_1 consists of 17 free nodes and 48 elements and is obtained by, first, a decomposition of Ω in 12 congruent squares of size $1/2$ and, second, a decomposition of each of the boxes along its two diagonals into 4 congruent triangles. The subsequent meshes are successively red-refined (i.e. each triangle is partitioned into 4 congruent subtriangles of Figure 15, left). This defines the (conforming) P_1 finite element spaces $V_1 \subset V_2 \subset V_3 \subset \dots \subset V := H_0^1(\Omega)$. The error $e := u - u_j$ of the finite element solution $u_j = \tilde{u}$ in $V_j = \bar{V}$ of dimension $N = \dim(V_j)$ is measured in the energy norm (the Sobolev seminorm in $H^1(\Omega)$)

$$\begin{aligned} |e|_{1,2} &:= |e|_{H^1(\Omega)} := \left(\int_{\Omega} |De|^2 \, dx \right)^{1/2} = a(u - \tilde{u}, u + \tilde{u})^{1/2} \\ &= \left(|u|_{H^1(\Omega)}^2 - |\tilde{u}|_{H^1(\Omega)}^2 \right)^{1/2} \end{aligned}$$



Notice that the two axes in Figure 11 scale logarithmically such that any algebraic curve of growth α is mapped into a straight line of slope $-\alpha$. The experimental convergence rate is $2/3$ in agreement with the (generic) singularity of the domain and resulting theoretical predictions. More details can be found in Carstensen, Bartels and Klose (2001).

For the benchmark problem of Section 5.6.1, the error estimator (75) can be written in the form

$$\eta_{R,R} := \left(\sum_{T \in \mathcal{T}} h_T^2 \|1\|_{L_2(T)}^2 \right)^{1/2} + \left(\sum_{E \in \mathcal{E}_R} h_E \int_E \left[\frac{\partial \bar{u}}{\partial \nu_E} \right]^2 d\mathbf{s} \right)^{1/2} \quad (116)$$

This experimental evidence supports the design of more elaborate estimators: The stopping criterion (72) with the reliable explicit estimators may appear very cheap and easy. But the decision (72) may have too costly consequences.

For comparison, the two implicit estimators η_L and η_{EQ} are displayed in Figure 11 as functions of N . It is stressed that both estimators are efficient and reliable (Carstensen and Funken, 1999/00)

$$|e|_{1,2} \leq \eta_r \leq 2.37 |e|_{1,2}$$

The practical performance of η_L and η_{EQ} in Figure 11 is comparable and in fact is much sharper than that of $\eta_{R,E}$ and $\eta_{R,R}$.

The averaging estimators η_A and η_M are as well displayed in Figure 11 as a function of N . Here, η_M is efficient up to higher-order terms (since the exact solution $u \in H^{5/3-\epsilon}(\Omega)$ is singular, this is not really guaranteed) while its reliability is open, i.e. the corresponding constants have not been computed. Nevertheless, the behavior of η_A and η_M is exclusively seen here from an experimental point of view.

The benchmark in Figure 11 is based on a sequence of uniform meshes and hence results in an experimental convergence rate 2/3 according to the corner singularity of this example. Adaptive mesh-refining algorithms, described below in more detail, are empirically studied also in Carstensen, Bartels and Klose (2001). The observations can be summarized as follows: The quality of the estimators and their relative accuracy is similar to what is displayed in Figure 11 even though the convergence rates are optimally improved to one.

This section provides a brief introduction to goal-oriented error control.

Given the Sobolev space $V = H_0^1(\Omega)$ with a finite-dimensional subspace $\tilde{V} \subset V$, a bounded and V -elliptic bilinear form $a : V \times V \rightarrow \mathbb{R}$, a bounded linear form $F : V \rightarrow \mathbb{R}$ there exists an exact solution $u \in V$ and a discrete solution $\tilde{u} \in \tilde{V}$ of

$$a(u, v) = F(v) \quad \forall v \in V \quad \text{and} \quad a(\tilde{u}, \tilde{v}) = F(\tilde{v}) \quad \forall \tilde{v} \in \tilde{V} \quad (117)$$

The previous sections concern estimations of the error $\epsilon := u - \tilde{u}$ in the energy norm, equivalent to the Sobolev norm $H^1(\Omega)$. In other norms are certainly of some interest as well as the error with respect to a certain goal functional. The latter is some given bounded and linear functional $J: V \rightarrow \mathbb{R}$ with respect to which one aims to monitor the error, that is, one wants to find computable lower and upper bounds for the (unknown) quantity

$$|J(u) - J(\bar{u})| = |J(e)|$$

Typical examples of goal functionals are described by L functions, for example,

$$J(v) = \int_{\Omega} qv \, dx \quad \forall v \in V$$

for a given $\rho \in L_2(\Omega)$ or as contour integrals

$$a(v, z) = J(v) \quad \forall v \in V \quad (118)$$

with exact solution $z \in V$ (guaranteed by the Lax–Milgram lemma) and the discrete solution $\tilde{z} \in \tilde{V}$ of

$$a(\tilde{v}, \tilde{z}) = J(\tilde{v}) \quad \forall \tilde{v} \in \tilde{V}$$

Set $f := z - \bar{z}$. On the basis of the Galerkin orthogonality $a(e, \bar{z}) = 0$ one infers

$$J(e) = a(e, z) = a(e, z - \bar{z}) = a(e, f) \quad (119)$$

As a result of (119) and the boundedness of a one obtains the a posteriori estimate

$$|J(e)| \leq \|a\| \|e\|_V \|f\|_V \leq \|a\| \eta_u \eta_v$$

Indeed, utilizing the primal and dual residual R_u and R_z in V^* , defined by

$$R_u := F - a(\tilde{u}, \cdot) \quad \text{and} \quad R_v := J - a(\cdot, \tilde{v})$$

computable upper error bounds for $\|e\|_V \leq \eta_u$ and $\|f\|_V \leq \eta_f$ can be found by the arguments of the energy error estimators of the previous sections. This yields a computable upper error bound $\|a\|_{\eta_u, \eta_f}$ for $|J(e)|$ which is global, that is, the interaction of e and f is not reflected. One might therefore speculate that the upper bound is often too coarse and inappropriate for goal-oriented adaptive mesh refinement.

Throughout the rest of this section, let the bilinear form a be symmetric and positive definite; hence a scalar product with induced norm $\|\cdot\|_a$. Then, the parallelogram rule shows

$$2J(e) = 2a(e, f) = \|e + f\|_a^2 - \|e\|_a^2 - \|f\|_a^2$$

This right-hand side can be written in terms of residuals in the spirit of (70), namely, $\|e\|_a = \|\text{Res}_u\|_{V^*}$, $\|f\|_a = \|\text{Res}_v\|_{V^*}$, and

$$\begin{aligned}\|e - f\|_a &= \|\text{Res}_{u+z}\|_V, \text{ for } \text{Res}_{u+z} := F + J - a(\tilde{u} + \tilde{z}), \\ &= \text{Res}_u + \text{Res}_z \in V^*\end{aligned}$$

Therefore, the estimation of $J(e)$ is reduced to the computation of lower and upper error bounds for the three residuals Res_s , Res_ϵ , and $\text{Res}_{s+\epsilon}$ with respect to the energy norm. This illustrates that the energy error estimation techniques of the previous sections may be employed for goal-oriented error control.

For more details and examples of a refined estimation see Ainsworth and Oden (2000) and Babeška and Strouboulis (2001).

5.7.4 Computing an approximation to $J(e)$

An immediate consequence of (119) is

$$J(e) = R(z)$$

and hence $J(e)$ is easily computed once the dual solution z of (118) is known or at least approximated to sufficient accuracy. An upper error bound for $|J(e)| = |R(z)|$ is obtained following the methodology of Becker and Rannacher (1996, 2001) and Bangerth and Rannacher (2003).

To outline this methodology, consider the residual representation formula (73) following the notation of Section 5.2.1. Suppose that $z \in H^2(\Omega)$ (e.g. for a H^2 regular dual problem) and let Iz denote the nodal interpolant in the lowest-order finite element space \tilde{V} . With some interpolation constant $C_I > 0$, there holds, for any element $T \in \mathcal{T}$,

$$h_T^2 \|z - Iz\|_{L^2(T)} + h_T^{3/2} \|z - Iz\|_{L^2(\partial T)} \leq C_I |z|_{H^2(T)}$$

The combination of this with (73) shows

$$\begin{aligned} J(e) &= \text{Res}(z - Iz) \\ &= \sum_{T \in \mathcal{T}} \int_T r_T \cdot (z - Iz) \, dx - \sum_{E \in \mathcal{E}} \int_E r_E \cdot (z - Iz) \, ds \\ &\leq \sum_{T \in \mathcal{T}} (\|r_T\|_{L^2(T)} \|z - Iz\|_{L^2(T)} \\ &\quad + \|r_E\|_{L^2(\partial T)} \|z - Iz\|_{L^2(\partial T)}) \\ &\leq \sum_{T \in \mathcal{T}} C_I \left(h_T^2 \|r_T\|_{L^2(T)} + h_T^{3/2} \|r_E\|_{L^2(\partial T)} \right) |z|_{H^2(T)} \end{aligned}$$

The influence of the goal functional in this upper bound is through the unknown H^2 seminorm $|z|_{H^2(T)}$, which is to be replaced by some discrete analog based on a computed approximation z_h . The justification of some substitute $|D_h^{k,h}|_{L^2(T)}$ (postprocessed with some averaging technique) for $|z|_{H^2(T)}$ is through striking numerical evidence; we refer to Becker and Rannacher (2001) and Bangerth and Rannacher (2003) for details and numerical experiments.

6 LOCAL MESH REFINEMENT

This section is devoted to the mesh-design task in the finite element method based on a priori and a posteriori information. Examples of the former type are graded meshes or geometric meshes with an a priori choice of refinement toward corner singularities briefly mentioned in Section 6.1. Examples of the latter type are adaptive algorithms for automatic mesh refinement (or mesh coarsening) strategies with a successive call of the steps

SOLVE \Rightarrow ESTIMATE \Rightarrow MARK \Rightarrow REFIN/COARSEN

Given the current triangulation, one has to compute the finite element solution in step SOLVE; cf. Section 7.8 for a MATLAB realization of that. The accuracy of this finite element approximation is checked in the step ESTIMATE. On the basis of the refinement indicators of Section 6.2 the step MARK identifies the elements, edges or patches in the current mesh in need of refinement (or coarsening). The new data structure is generated in step REFIN/COARSEN where a partition is given and a closure algorithm computes a triangulation described in Section 6.3. The convergence and optimality of the adaptive algorithm is discussed in Section 6.4. Brief remarks on coarsening strategies in Section 6.5 conclude the section.

6.1 A priori mesh design

The singularities of the exact solution of the Laplace equation on domains with corners (cf. Figure 1) are reasonably well understood and motivate the (possibly anisotropic) mesh refinement toward vertices or edges. This section aims a short introduction for two-dimensional P_1 finite elements – Chapter 3, this Volume, will report on three-dimensional examples.

Given a polygonal domain with a coarse triangulation into triangles (which specify the geometry), macro elements can be used to fill the domain with graded meshes. Figure 12(a) displays a macro element described in the sequel while Figure 12(b) illustrates the resulting fine mesh for an L-shaped domain.

The description is restricted to the geometry on the reference element T_{ref} with vertices $(0, 0)$, $(1, 0)$, and $(0, 1)$ of Figure 12(a). The general situation is then obtained by an affine transformation illustrated in Figure 12(b). The macro element T_{ref} is generated as follows: Given a grading parameter $\beta > 0$ for a grading function $g(t) = t^\beta$, and given a natural number N , set $\xi_j := g(j/N)$ and draw line segments aligned to the antidiagonal through $(0, \xi_j)$ and $(\xi_j, 0)$ for $j = 0, 1, \dots, N$. Each of these segments

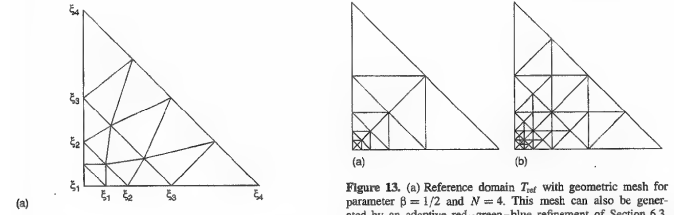


Figure 13. (a) Reference domain T_{ref} with geometric mesh for parameter $\beta = 1/2$ and $N = 4$. This mesh can also be generated by an adaptive red-green-blue refinement of Section 6.3. (b) Illustration of the closure algorithm. The refinement triangulation with 50 element domains is obtained from the mesh (a) with 18 element domains by marking one edge (namely the second along the antidiagonal) in the mesh (a).

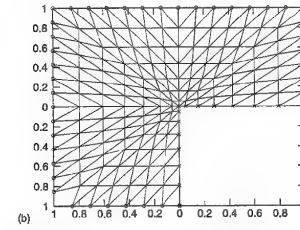


Figure 12. (a) Reference domain T_{ref} with graded mesh for $\beta = 3/2$ and $N = 4$. (b) Graded mesh on L-shaped domain with refinement toward origin and uniform refinement far away from the origin. Notice that the outer boundaries of the macro elements show a uniform distribution and so match each other in one global regular triangulation.

is divided into j uniform edges and so define the set of nodes $(0, 0)$ and ξ_j/j ($j - k, k$) for $k = 0, \dots, j$ and $j = 1, \dots, N$. The elements are then given by the vertices ξ_j/j ($j - k, k$) and $\xi_{j+1}/(j + 1)$ ($j - k - 1, k$) on the antidiagonal and the vertex $\xi_{j+1}/(j + 1)$ ($j - k - 1, k$) on the finer and $\xi_{j+1}/(j + 1)$ ($j - k, k + 1$) on the coarser neighboring segment, respectively. The finest element is $\text{conv}((0, 0), (0, \xi_1), (\xi_1, 0))$ of diameter $\sqrt{(2)} g(1/N) \approx N^{-\beta}$.

Figure 12(b) displays a triangulation of the L-shaped domain with a refinement toward the origin designed by a union of transformed macro elements with $\beta = 3/2$ and $N = 7$. The other vertices of the L-shaped domain yield higher singularities, which are not important for the first-order Courant finite element.

The geometric mesh depicted in Figure 13 yields a finer refinement toward some corners of the polygonal domain. Given a parameter $\beta > 0$ in this type of triangulation,

the nodes $\xi_0 := 0$ and $\xi_j := \beta^{N-j}$ for $j = 1, \dots, N$ define antidiagonals through $(\xi_j, 0)$ and $(0, \xi_j)$, which are in turn bisected. For such a triangulation, the polynomial degrees p_T on each triangle T are distributed as follows: $p_T = 1$ for the two triangles T with vertex $(0, 0)$ and $p_T = j + 2$ for the four elements in the convex quadrilateral with the vertices $(\xi_j, 0)$, $(0, \xi_j)$, $(\xi_{j+1}, 0)$, and $(0, \xi_{j+1})$ for $j = 0, \dots, N - 1$. Figure 14 compares experimental convergence rates of the error in H^{-1} -seminorm $|e|_{H^{-1}}$ for various graded meshes for the P_1 finite element method, the p - and hp -finite element method, and for the adaptive algorithm of Section 6.4. The P_1 finite element method on graded meshes with $\beta = 3/2$, $\beta = 2$ and h -adaptivity recover optimality in the convergence rate as opposite to the uniform refinement ($\beta = 1$), leading only to a sub-optimal convergence due to the corner singularity. The hp -finite element method performs better convergence rate compared to the p -finite element method.

Tensor product meshes are more appropriate for smaller values of β ; the one-dimensional model analysis of Babeška and Guo (1986) suggests $\beta = (2)^{1/2} - 1 \approx 0.171573$.

6.2 Adaptive mesh-refining algorithms

The automatic mesh refinement for regular triangulations called MARK and REFIN/COARSEN frequently consists of three stages:

- (i) the marking of elements or edges for refinement (or coarsening);
- (ii) the closure algorithm to ensure that the resulting triangulation is (or remains) regular;
- (iii) the refinement itself, i.e. the change of the underlying data structures.

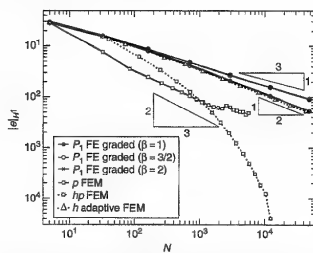


Figure 14. Experimental convergence rates for various graded meshes for the P_1 finite element method, the p - and hp -finite element method, and for the adaptive algorithm of Section 6.4 for the Poisson problem on the L -shaped domain.

This section will focus on the marking strategies (i) while the subsequent section addresses the closure algorithm (ii) and the refinement procedure (iii).

In a model situation with a sum over all elements $T \in \mathcal{T}$ (or over all edges, faces, or nodes), the a posteriori error estimators of the previous section give rise to a lower or upper error bound η in the form

$$\eta = \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{1/2}$$

Then, the marking strategy is an algorithm, which selects a subset \mathcal{M} of \mathcal{T} called the *marked* elements; these are marked with the intention of being refined during the later refinement algorithm.

A typical algorithm computes a threshold L , a positive real number, and then utilizes the *refinement rule* or *marking criterion*

$$\text{mark } T \in \mathcal{T} \text{ if } L \leq \eta_T$$

Therein, η_T is referred to as the *refinement indicator* whereas L is the *threshold*; that is,

$$\mathcal{M} := \{T \in \mathcal{T} : L \leq \eta_T\}$$

Typical examples for the computation of a threshold L are the *maximum criterion*

$$L := \Theta \max\{\eta_T : T \in \mathcal{T}\}$$

or the *bulk criterion* where L is the largest value such that

$$(1 - \Theta)^2 \sum_{T \in \mathcal{T}} \eta_T^2 \leq \sum_{T \in \mathcal{M}} \eta_T^2$$

The parameter Θ is chosen with $0 \leq \Theta \leq 1$; $\Theta = 0$ corresponds to an almost uniform refinement and $\Theta = 1$ to a raw refinement of just a few elements (no refinements in the bulk criterion).

The justification of the refinement criteria is essentially based on the heuristic of an equi-distribution of the refinement indicator; see Babuška and Rheinboldt (1978, 1979) and Babuška and Vogelius (1984) for results in 1-D. A rigorous justification for some class of problems started with Dörfler (1996); it is summarized in Morin, Nochetto and Siebert (2003b) and will be addressed in Section 6.4.

A different strategy is possible if the error estimator gives rise to a quantitative bound of a new meshsize. For instance, the explicit error estimator can be rewritten as $\eta_R := \|h_T R\|_{L_2(\Omega)}$ with a given function $R \in L_2(\Omega)$ and the local mesh size h_T (when edge contributions are recast as volume contributions). Then, given a tolerance Tol and using the heuristic that R would not change (at least not dramatically increase) during a refinement, the new local mesh size h^{new} can be calculated from the condition

$$\|h^{\text{new}} R\|_{L_2(\Omega)} = \text{Tol}$$

upon the equi-distribution hypothesis $h^{\text{new}} \propto \text{Tol}/R$. Another approach that leads to a requested mesh-size distribution is based on sharp a priori error bounds, such as $\|h_T D^2 u\|_{L_2(\Omega)}$ where $D^2 u$ denotes the matrix of all second derivatives of the exact solution u . Since $D^2 u$ is unknown, it has to be approximated by postprocessing a finite element solution with some averaging technique.

The aforementioned refinement rules for the step MARK (intended for conforming FEM) ignore further decisions such as the particular type of anisotropic refinement or the increase of polynomial degrees versus the mesh refinements for hp -FEM. One way to handle such subtle decisions will be addressed under the heading coarsening strategy in Section 6.5.

6.3 Mesh refining of regular triangulations

Given a marked set of objects such as nodes, edges, faces, or elements, the refinement of the element (triangles or tetrahedra) plus further refinements (closure algorithm) for the design of regular triangulations are considered in this section.



Figure 15. Red-, green-, blue-(left)- and blue-(right) refinement with reference edge on bottom of a triangle (from left to right) into four, two, and three subtriangles. The bold lines opposite the newest vertex indicate the next reference edge for further refinements.

6.3.1 Refinement of a triangle

Triangular elements in two dimensions are refined into two, three, or four subtriangles as indicated in Figure 15. All these divisions are based on hidden information on some reference edge. Rivara (1984) assumed the longest edge in the triangle as base of the refinement strategies while the one below is based on the explicit marking of a reference edge. In Figure 15, the bottom edge of the original triangle acts as the *reference edge* and, in any refinement, is halved. The four divisions displayed correspond to the bisection of one (called green-refinement), of two (the two versions of blue-refinement), or of all three edges (for red-refinement)

as long as the bottom edge is refined. In Figure 15, the reference edges in the generated subtriangles are drawn with a bold line.

6.3.2 Refinement of a tetrahedron

The three-dimensional situation is geometrically more complicated. Therefore, this section is focused on the bisection algorithm and the readers are referred to Bey (1995) for 3-D red-refinements.

Figure 16 displays five types of tetrahedra which shows the bottom triangle with vertices of local label 1, 2, and

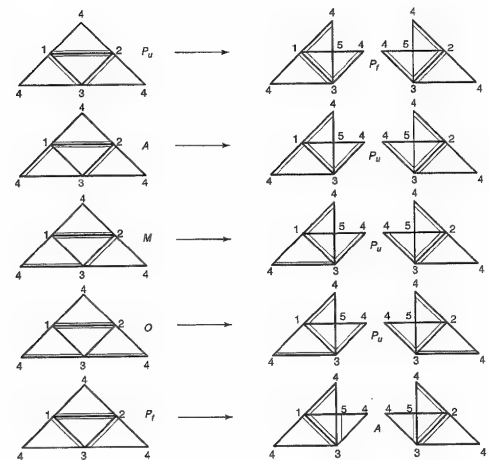


Figure 16. Bisection rules in 3-D after Arnold, Mukherjee and Pouly (2000) which essentially dates back to Bänsch (1991). It involves five types of tetrahedra (indicated as P_u , A , M , O , and P_f on the left) and their bisections with an initialization of $M \rightarrow 2 P_u$ and $O \rightarrow 2 P_u$, followed by a successive loop of the form $A \rightarrow 2 P_u \rightarrow 4 P_f \rightarrow 8 A$ and so on. Notice that the forms of P_u and P_f are identical, but play a different role in the refinement loop.

3 while the remaining three faces with the vertex label 4 are drawn in the same plane as the bottom face; hence, the same number 4 for the top vertex is visible at the end-points of the three outer subtriangles.

Each triangular face has a reference edge and (at least) one edge is a reference edge of the two neighboring faces and that determines the vertices with local numbers 1 and 2. In the language of Arnold, Mukherjee and Pouly (2000), the five appearing types are called P_1 , A , M , O , and P_2 . The bisection of the edges between the vertices number 1 and 2 with the new vertex number 5 are indicated in Figure 16.

It is an important property of the inherited reference edges that the edges of each element K in the original regular triangulation \mathcal{T} are refined in a cyclic way such that four consecutive bisections inside K are necessary before a new vertex is generated, which is not the midpoint of some edge of K .

6.3.3 Closure algorithms

The bisection of some set of marked elements does not always lead to a regular triangulation – the new vertices may be hanging nodes for the neighboring elements. Further refinements are necessary to make those nodes regular. The default procedure within the class of bisection algorithms is to work on a set of marked elements $\mathcal{M}^{(k)}$.

Closure Algorithm for Bisection. Input a regular triangulation $\mathcal{T}^{(0)} := \mathcal{T}$ and an initial subset $\mathcal{M}^{(0)} := \mathcal{M} \subset \mathcal{T}^{(0)}$ of marked elements, set $Z = \emptyset$ and $k := 0$.

While $\mathcal{M}^{(k)} \neq \emptyset$ repeat (i)–(iv):

- (i) choose some element K in $\mathcal{M}^{(k)}$ with reference edge E and initiate its midpoint z_E as new vertex, set $Z := Z \cup \{z_E\}$;
- (ii) bisect E and divide K into K_+ and K_- and set $\mathcal{T}^{(k+1)} := \{K_+, K_-\} \cup (\mathcal{T}^{(k)} \setminus \{K\})$;
- (iii) find all elements $T_1, \dots, T_{m_E} \in \mathcal{T}^{(k+1)}$ with hanging node $z \in Z$ (if any) and set $\mathcal{M}^{(k+1)} := \{T_1, \dots, T_{m_E}\} \cup (\mathcal{M}^{(k)} \setminus \{K\})$;
- (iv) update $k := k + 1$ and go to (i).

Output a refined triangulation $\mathcal{T}^{(k)}$.

According to step (iii) of the closure algorithm, any termination leads to a regular triangulation $\mathcal{T}^{(k)}$. The remaining essential detail is to guarantee that there will always occur a termination via $\mathcal{M}^{(k)} = \emptyset$ for some k . In the newest-vertex bisection, for instance, the reference edges are inherited in such a way that any element $K \in \mathcal{T}$, the initial regular triangulation, is refined only by subdivisions which, at most, halve each edge of K . Since the closure algorithm only halves some edge in \mathcal{T} and prohibits any further refinements,

any intermediate (irregular) $\mathcal{T}^{(k)}$ remains coarser than or equal to some regular uniform refinement $\widehat{\mathcal{T}}$ of \mathcal{T} . This is the main argument to prove that the closure algorithm cannot refine forever and stops after a finite number of steps.

Figure 13(b) shows an example where, given the initial mesh of Figure 13(a), only one edge, namely the second on the diagonal, is marked for refinement and the remaining refinement is induced by the closure algorithm. Nevertheless, the number of new elements can be bounded in terms of the initial triangulation and the number of marked elements (Binev, Dahmen and DeVore, 2004).

The closure algorithm for the red–green–blue refinement in 2-D is simpler when the focus is on marking of edges. One main ingredient is that each triangle K is assigned a reference edge $E(K)$. If we are given a set of marked elements, let \mathcal{M} denote the set of corresponding assigned reference edges.

Closure Algorithm for Red–Green–Blue Refinement. Input a regular triangulation \mathcal{T} with a set of edges \mathcal{E} and an initial subset $\mathcal{M} \subset \mathcal{E}$ of marked edges, set $k := 0$ and $\mathcal{M}^{(0)} := \mathcal{M}$.

While $\mathcal{M}^{(k)} \neq \emptyset$ repeat (i)–(iv):

- (i) choose some edge E in $\mathcal{M}^{(k)}$ and let $T_{\pm} \in \mathcal{T}$ denote the (at most) two triangles that share the edge $E \subset \partial T_{\pm}$;
- (ii) set $\mathcal{M}^{(k+1)} := \mathcal{M}^{(k)} \cup \{E_+, E_-\}$ with the reference edge $E_{\pm} := E(T_{\pm})$ of T_{\pm} ;
- (iii) if $\mathcal{M}^{(k+1)} := \mathcal{M}^{(k)}$ set $\mathcal{N}^{(k+1)} := \mathcal{N}^{(k)} \setminus \{E\}$ else set $\mathcal{N}^{(k+1)} := (\mathcal{N}^{(k)} \cup \{E_+, E_-\}) \setminus \{E\}$;
- (iv) update $k := k + 1$ and go to (i).

Bisect the marked edges $\{E \in \mathcal{M}^{(k)} : E \subset \partial T\}$ of each element $T \in \mathcal{T}$ and refine \mathcal{T} by one of the red–green–blue refinement rules to generate elementwise a partition $\widehat{\mathcal{T}}$.

Output the regular triangulation $\widehat{\mathcal{T}}$.

The closure algorithm for red–green–blue refinement terminates as $\mathcal{N}^{(k)}$ is decreasing and $\mathcal{M}^{(k)}$ is increasing and outputs a set $\widehat{\mathcal{M}} := \mathcal{M}^{(k)}$ of marked edges with the following closure property: Any element $T \in \mathcal{T}$ with an edge in $\widehat{\mathcal{M}}$ satisfies $E(T) \in \widehat{\mathcal{M}}$, i.e. if T is marked, then at least its reference edge will be halved. This property allows the application of one properly chosen refinement of Figure 15 and leads to a regular triangulation.

The reference edge $E(K)$ in the closure algorithm is assigned to each element K of the initial triangulation and then is inherited according to the rules of Figure 15. For newest-vertex bisection, each triangle with vertices of global numbers j, k , and ℓ has the reference edge opposite to the vertex number $\max\{j, k, \ell\}$.

On the basis of refinement rules that inherit a reference edge to the generated elements, one can prove that a finite number of affine-equivalent elements domains occur.

6.4 Convergence of adaptive algorithms

This section discusses the convergence of a class of adaptive P_1 finite element methods based on newest-vertex bisection or red–green–blue refinement of Section 6.3.1. For the ease of this presentation, let us adopt the notation of the residual representation formula (73) of Section 5.2.1 in 2-D and perform a refinement with four bisections of each triangle with a marked edge as illustrated in Figure 17.

Adaptive Algorithm. Input a regular triangulation $\mathcal{T}^{(0)} := \mathcal{T}$ with a set of reference edges $\{E(K) : K \in \mathcal{T}^{(0)}\}$ and a parameter $0 < \Theta \leq 1$, set $k := 0$.

Repeat (i)–(v) until termination:

- (i) compute discrete solution $u_k \in V_k$ based on the triangulation $\mathcal{T}^{(k)}$ with set of interior edges $\mathcal{E}^{(k)}$;
- (ii) compute $\eta_k^2 := h_E \|r_E\|_{L_2(E)}^2 + \sum_{T \in \mathcal{T}(E)} h_T^2 \|r_T\|_{L_2(T)}^2$ for any $E \in \mathcal{E}^{(k)}$ with the set $\mathcal{T}(E)$ of its at most two neighboring elements;
- (iii) mark edges in $\mathcal{M}^{(k)} \subset \mathcal{E}^{(k)}$ by the bulk criterion such that

$$(1 - \Theta)^2 \sum_{E \in \mathcal{E}^{(k)}} \eta_E^2 \leq \sum_{E \in \mathcal{M}^{(k)}} \eta_E^2$$

- (iv) for any $E \in \mathcal{M}^{(k)}$, bisect each element $T \in \mathcal{T}(E)$ (at least) four times (as in Figure 17(a)) such that (at least) each midpoint z_E of any edge $E \in \mathcal{E}(T)$ and the midpoint of T become new nodes (cf. Figure 17(b));
- (v) call closure algorithm to generate a refined regular triangulation $\mathcal{T}^{(k+1)}$, update $k := k + 1$ and go to (i).

Output a sequence of finite element meshes $\mathcal{T}^{(k)}$ with associated finite element spaces V_k and discrete solutions u_k . Recall that u is the exact solution. Then there holds the error-reduction property

$$\|u - u_{k+1}\|_a^2 \leq \Theta \|u - u_k\|_a^2 + \text{osc}(f)^2$$

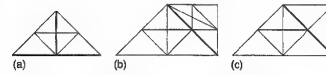


Figure 17. (a) Refinement of triangle with four bisections from the adaptive algorithm of Section 6.4. (b) Refinement of the edge neighborhood ω_E with four bisections of each element T_E . (c) Refinement of ω_E with four bisections of one element T_E and green-refinement of T_- .

for any k and with a constant $0 \leq \Theta < 1$, which depends on $\mathcal{T}^{(0)}$ and $\Theta > 0$. The oscillations in $\text{osc}(f)$ may be seen as higher-order terms when compared with the volume term $\|h_T f\|_{L_2(\Omega)}$ of the error estimator.

An outline of the proof of the error-reduction property concludes this section to illustrate that $\Theta < 1$ is independent of any mesh size, the level k , and number of elements $N_k := \text{card}(\mathcal{T}^{(k)})$. The proof follows the arguments of Section 5.4.3 for $V_H := \widehat{V} := V_k$ and $V_h := V_H \oplus W_h \subseteq V_{k+1}$. The space W_h is spanned by all b_E and by all b_T for all $T \in \mathcal{T}(E)$ for $E \in \mathcal{M}^{(k)} \subset \mathcal{E}^{(k)}$, which, here, are given by a nodal basis function in the new triangulation of the new interior node of $T \in \mathcal{T}(E)$ and of the midpoint of $E \in \mathcal{M}^{(k)}$. These basis functions substitute the bubble functions (80) in Section 5.2.4 and have in fact the properties (81)–(83). One then argues as in Section 5.4.3 with some

$$w_h := \sum_{E \in \mathcal{M}^{(k)}} (w_E + \sum_{T \in \mathcal{T}(E)} w_T) \in W_h$$

to show $\|u_h\|_a^2 \leq C(R(u_h) + \text{osc}(f))$ and eventually

$$\sum_{E \in \mathcal{M}^{(k)}} \eta_E^2 \leq C(\|u_h - u_H\|_a^2 + \text{osc}(f)^2)$$

for $u_h := u_{k+1}$ and $u_H := u_k$. This and the bulk criterion lead to

$$\begin{aligned} \|u - u_h\|_a^2 &\leq C_{\text{rel}} \left(\sum_{E \in \mathcal{E}^{(k)}} \eta_E^2 + \text{osc}(f)^2 \right) \\ &\leq (1 - \Theta)^{-2} C_{\text{rel}} \left(\sum_{E \in \mathcal{M}^{(k)}} \eta_E^2 + \text{osc}(f)^2 \right) \\ &\leq C (1 - \Theta)^{-2} C_{\text{rel}} (\|u_h - u_H\|_a^2 + \text{osc}(f)^2) \end{aligned}$$

Utilizing the Galerkin orthogonality

$$\|u_h - u_H\|_a^2 = \|u - u_H\|_a^2 - \|u - u_h\|_a^2$$

one concludes the proof.

It is interesting to notice that the conditions on the refinement can be weakened. For instance, it suffices to refine solely one element in $\mathcal{T}(E)$ such that there is an interior node as in Figure 17 (c). (But then the oscillations are coarser and involve terms like $h_E \|f - f_E\|_{L_2(\omega_E)}$ with an integral mean f_E over an edge-patch ω_E .) However, the counterexample of Figure 10 clearly indicates that some specific conditions are necessary to ensure the error-reduction property.

6.5 Coarsening

The constructive task in approximation theory is the design of effective approximations of a given function u . Replacing the unknown u by a very accurate known approximation (e.g. obtained by some overkill computation) of it, one may employ all the approximation techniques available and find an efficient representation in terms of an adapted finite element space, i.e. in terms of an underlying adapted mesh and/or an adapted distribution of polynomial degrees.

The input of this coarsening step consists, for instance, of a given triangulation T_k and a given finite element solution u_k and some much more accurate approximation \tilde{u}_k . Then, the difference

$$\eta_k := \|u_k - \tilde{u}_k\|_a$$

is an error estimator as accurate as $\|u - u_k\|_a$ (reliability and efficiency is proved by a triangle inequality). Moreover, one might use $\eta_T = |\tilde{u}_k - u_k|_{H^1(T)}$ as a local refinement indicator in the setting of the benchmark example.

Moreover, many decisions of the mesh design (e.g. choice of anisotropic elements, choice of polynomial degree, etc.) can be based on the explicit availability of \tilde{u}_k .

One single coarsening step is certainly pointless, for the reference solution \tilde{u}_k is known and regarded as very accurate (and so there is no demand for further work). On the other hand, a cascade of coarsening steps requires in each step, a high accuracy relative to the precision of that level and hence a restricted accuracy. A schematic description of this procedure reads as follows:

Coarsening Algorithm for Adaptive Mesh Design. Input a finite element space V_0 with a finite element solution u_0 , set $k := 0$.

Repeat (i)–(iv) until termination:

- (i) design a super space \tilde{V}_k of V_k by uniform refinements plus uniform increase of all the polynomial degrees;
- (ii) compute accurate finite element solution \tilde{u}_k with respect to \tilde{V}_k ;
- (iii) coarsen \tilde{V}_k based on given \tilde{u}_k and design a new finite element space V_{k+1} ;
- (iv) update $k := k + 1$ and go to (i).

The paper of Binev, Dahmen and DeVore (2004) presents a realization of this algorithm with an adaptive algorithm on stage (i) and then proves that the convergence rate of the overall procedure is optimal with respect to the energy error as a function of the number of degrees of freedom.

An automatic version of the coarsening strategy is called the *hp-adaptive finite element method* in Demkowicz (2003)

for higher-order polynomials adaptivity. This involves algorithms for the determination of edges that will be refined and of their polynomial degrees.

7 OTHER ASPECTS

In this section, we discuss briefly several topics in finite element methodology. Some of the discussions involve the Sobolev space $W_p^k(\Omega)$ ($1 \leq p \leq \infty$), which is the space of functions in $L_p(\Omega)$ whose weak derivatives up to order k also belong to $L_p(\Omega)$, with the norm

$$\|v\|_{W_p^k(\Omega)} = \left(\sum_{|\alpha| \leq k} \left\| \frac{\partial^\alpha v}{\partial x^\alpha} \right\|_{L_p(\Omega)} \right)^{1/p}$$

for $1 \leq p < \infty$ and

$$\|v\|_{W_\infty^k(\Omega)} = \max_{|\alpha| \leq k} \left\| \frac{\partial^\alpha v}{\partial x^\alpha} \right\|_{L_\infty(\Omega)}$$

For $1 \leq p < \infty$, the seminorm $(\sum_{|\alpha|=k} \|\partial^\alpha v / \partial x^\alpha\|_{L_p(\Omega)})^{1/p}$ will be denoted by $|v|_{W_p^k(\Omega)}$, and the seminorm $\max_{|\alpha|=k} \|\partial^\alpha v / \partial x^\alpha\|_{L_\infty(\Omega)}$ will be denoted by $|v|_{W_\infty^k(\Omega)}$.

7.1 Nonsymmetric/indefinite problems

The results in Section 4 can be extended to the case where the bilinear form $a(\cdot, \cdot)$ in the weak problem (1) is nonsymmetric and/or indefinite due to lower order terms in the partial differential equation. We assume that $a(\cdot, \cdot)$ is bounded (cf. (2)) on the closed subspace V of the Sobolev space $H^m(\Omega)$ and replace (3) by the condition that

$$a(v, v) + L\|v\|_{L_2(\Omega)}^2 \geq C_3\|v\|_{H^m(\Omega)}^2 \quad \forall v \in V \quad (120)$$

where L is a positive constant.

Example 15. Let $a(\cdot, \cdot)$ be defined by

$$\begin{aligned} a(v_1, v_2) = & \int_\Omega \nabla v_1 \cdot \nabla v_2 \, dx + \sum_{j=1}^d \int_\Omega b_j(x) \frac{\partial v_1}{\partial x_j} v_2 \, dx \\ & + \int_\Omega c(x) v_1 v_2 \, dx \end{aligned} \quad (121)$$

for all $v_1, v_2 \in H^1(\Omega)$, where $b_j(x)$ ($1 \leq j \leq d$), $c(x) \in L_\infty(\Omega)$. If we take $V = \{v \in H^1(\Omega) : v|_\Gamma = 0\}$ and F is

defined by (6), then (1) is the weak form of the nonsymmetric boundary value problem

$$\begin{aligned} -\Delta u + \sum_{j=1}^d b_j \frac{\partial u}{\partial x_j} + cu = f, \quad u = 0 \quad \text{on } \Gamma, \\ \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega \setminus \Gamma \end{aligned} \quad (122)$$

and the coercivity condition (120) follows from the well-known Gårding's inequality (Agmon, 1965).

Unlike the symmetric positive definite case, we need to assume that the weak problem (1) has a unique solution, and that the adjoint problem is also uniquely solvable, that is, given any $G \in V^*$ there is a unique $w \in V$ such that

$$a(v, w) = G(v) \quad \forall v \in V \quad (123)$$

Furthermore, we assume that the solution w of (123) enjoys some elliptic regularity when $G(v) = (g, v)_{L_2(\Omega)}$ for $g \in L_2(\Omega)$, i.e., $w \in H^{m+\alpha}(\Omega)$ for some $\alpha > 0$ and

$$\|w\|_{H^{m+\alpha}(\Omega)} \leq C\|g\|_{L_2(\Omega)} \quad (124)$$

Let T be a triangulation of Ω with mesh size $h_T = \max_{T \in \mathcal{T}} \text{diam } T$ and $V_T \subset V$ be a finite element space associated with T such that the following approximation property is satisfied:

$$\inf_{v \in V_T} \|w - v\|_{H^m(\Omega)} \leq \epsilon_T \|w\|_{H^{m+\alpha}(\Omega)} \quad \forall w \in H^{m+\alpha}(\Omega) \quad (125)$$

where

$$\epsilon_T \downarrow 0 \quad \text{as } h_T \downarrow 0 \quad (126)$$

The discrete problem is then given by (54).

Following Schatz (1974) the well-posedness of the discrete problem and the error estimate for the finite element approximate solution can be addressed simultaneously. Assume for the moment that $u_T \in V_T$ is a solution of (54). Then we have

$$a(u - u_T, v) = 0 \quad \forall v \in V_T \quad (127)$$

We use (127) and a duality argument to estimate $\|u - u_T\|_{L_2(\Omega)}$ in terms of $\|u - u_T\|_{H^m(\Omega)}$. Let $w \in V$ satisfy

$$a(v, w) = (u - u_T, v)_{L_2(\Omega)} \quad \forall v \in V_T \quad (128)$$

We obtain, from (2), (127), and (128), the following analog of (24):

$$\|u - u_T\|_{L_2(\Omega)}^2 = a(u - u_T, w) \leq C \left(\inf_{v \in V_T} \|w - v\|_{H^m(\Omega)} \right)$$

$$\times \|u - u_T\|_{H^m(\Omega)} \quad (129)$$

and hence, by (124) and (125),

$$\|u - u_T\|_{L_2(\Omega)} \leq \epsilon_T \|u - u_T\|_{H^m(\Omega)} \quad (130)$$

It follows from (120) and (130) that

$$\|u - u_T\|_{H^m(\Omega)} \leq a(u - u_T, u - u_T) + C\epsilon_T^2 \|u - u_T\|_{H^m(\Omega)}^2$$

which together with (126) implies, for h_T sufficiently small,

$$\|u - u_T\|_{H^m(\Omega)}^2 \leq a(u - u_T, u - u_T) \quad (131)$$

For the special case where $F = 0$ and $u = 0$, any solution u_T of the homogeneous discrete problem

$$a(u_T, v) = 0 \quad \forall v \in V_T$$

must satisfy, by (131),

$$\|u_T\|_{H^m(\Omega)}^2 \leq 0$$

We conclude that any solution of the homogeneous discrete problem must be trivial and hence the discrete problem (19) is uniquely solvable provided h_T is sufficiently small. Under this condition, we also obtain immediately from (2), (127), and (131), the following analog of (22):

$$\|u - u_T\|_{H^m(\Omega)} \leq C \inf_{v \in V_T} \|u - v\|_{H^m(\Omega)} \quad (132)$$

Concrete error estimates now follow from (132), (129), and the results in Section 3.3.

7.2 Nonconforming finite elements

When the finite element space FE_T defined by (32) does not belong to the Sobolev space $H^m(\Omega)$ where the weak problem (1) of the boundary value problem is posed, it is referred to as a *nonconforming* finite element space. Nonconforming finite element spaces are more flexible and are useful for problems with constraints where conforming finite element spaces are more difficult to construct.

Example 16 (Triangular Nonconforming Elements) Let K be a triangle. If the set \mathcal{N}_K consists of evaluations of the shape functions at the midpoints of the edges of K (Figure 18a), then (K, P_1, \mathcal{N}_K) is the nonconforming P_1 element of Crouzeix and Raviart. It is the simplest triangular element that can be used to solve the incompressible Stokes equation.

If the set N_K consists of evaluations of the shape functions at the vertices of K and the evaluations of the normal derivatives of the shape functions at the midpoints of the edges of K (Figure 18b), then (K, P_2, N_K) is the Morley element. It is the simplest triangular element that can be used to solve the plate bending problem.

Example 17 (Rectangular Nonconforming Elements)

Let K be a rectangle. If P_K is the space spanned by the functions $1, x_1, x_2$ and $x_1^2 - x_2^2$ and the set N_K consists of the mean values of the shape functions on the edges of K , then (K, P_K, N_K) is the rotated Q_1 element of Rannacher and Turek (Figure 19a), where the thick lines represent mean values over the edges). It is the simplest rectangular element that can be used to solve the incompressible Stokes equation.

If P_K is the space spanned by the functions $1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^2x_2$ and $x_1x_2^2$ and the set N_K consists of evaluations of the shape functions at the vertices of K and evaluations of the normal derivatives at the midpoints of the edges (Figure 19b), then (K, P_K, N_K) is the incomplete Q_2 element. It is the simplest rectangular element that can be used to solve the plate bending problem.

Consider the weak problem (1) for a symmetric positive definite boundary value problem, where F is defined by (6) for a function $f \in L_2(\Omega)$. Let V_T be a nonconforming finite element space associated with the triangulation T . We assume that there is a (mesh-dependent) symmetric bilinear

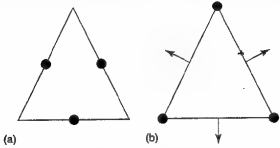


Figure 18. Triangular nonconforming finite elements.

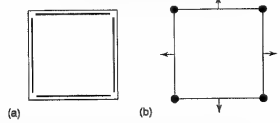


Figure 19. Rectangular nonconforming finite elements.

form $a_T(\cdot, \cdot)$ defined on $V + V_T$ such that (i) $a_T(v, v) = a(v, v)$ for $v \in V$, (ii) $a_T(\cdot, \cdot)$ is positive definite on V_T . The discrete problem for the nonconforming finite element method reads: Find $u_T \in V_T$ such that

$$a_T(u_T, v) = \int_{\Omega} f v \, dx \quad \forall v \in V_T \quad (133)$$

Example 18. The Poisson problem in Example 1 can be solved by the nonconforming P_1 finite element method in which the finite element space is $V_T = \{v \in L_2(\Omega) : v|_T \in P_1(T) \text{ for every triangle } T \in T, v \text{ is continuous at the midpoints of the edges of } T \text{ and } v \text{ vanishes at the midpoints of the edges of } T \text{ along } \Gamma\}$ and the bilinear form a_T is defined by

$$a_T(v_1, v_2) = \sum_{T \in T} \int_T \nabla v_1 \cdot \nabla v_2 \, dx \quad (134)$$

The nonconforming Ritz-Galerkin method (133) can be analyzed as follows. Let $\tilde{u}_T \in V_T$ be defined by

$$a_T(\tilde{u}_T, v) = a_T(u, v) \quad \forall v \in V_T$$

Then we have

$$\|u - \tilde{u}_T\|_{a_T} = \inf_{v \in V_T} \|u - v\|_{a_T}$$

where $\|u\|_{a_T} = (a_T(u, u))^{1/2}$ is the nonconforming energy norm defined on $V + V_T$, and we arrive at the following generalization (Berger, Scott and Strang, 1972) of (21):

$$\begin{aligned} \|u - u_T\|_{a_T} &\leq \|u - \tilde{u}_T\|_{a_T} + \|\tilde{u}_T - u_T\|_{a_T} \\ &= \inf_{v \in V_T} \|u - v\|_{a_T} + \sup_{v \in V_T(0)} \frac{a_T(\tilde{u}_T - u_T, v)}{\|v\|_{a_T}} \\ &= \inf_{v \in V_T} \|u - v\|_{a_T} + \sup_{v \in V_T(0)} \frac{a_T(u - u_T, v)}{\|v\|_{a_T}} \end{aligned} \quad (135)$$

Remark 17. The second term on the right-hand side of (135), which vanishes in the case of conforming Ritz-Galerkin methods, measures the consistency errors of nonconforming methods.

As an example, we analyze the nonconforming P_1 method for the Poisson problem in Example 18. For each $T \in T$, we define an interpolation operator $\Pi_T : H^1(T) \rightarrow P_1(T)$ by

$$(\Pi_T \zeta)(m_e) = \frac{1}{|e|} \int_e \zeta \, ds$$

where m_e is the midpoint for the edge e of T . The interpolation operator Π_T satisfies the estimate (43) for $m = 1$, and they can be pieced together to form an interpolation operator $\Pi : H^1(\Omega) \rightarrow V_T$. Since the solution u of (5) belongs to $H^{1+\alpha(T)}(T)$, where $0 < \alpha(T) \leq 1$, the first term on the right-hand side of (135) satisfies the estimate

$$\inf_{v \in V_T} \|u - v\|_{a_T} \leq \|u - \Pi u\|_{a_T} \leq C \left(\sum_{T \in T} (\text{diam } T)^{2\alpha(T)} |u|_{H^{1+\alpha(T)}(T)}^2 \right)^{1/2} \quad (136)$$

where the constant C depends only on the minimum angle in T .

To analyze the second term on the right-hand side of (135), we write, using (5), (133), and (134),

$$a_T(u - u_T, v) = - \sum_{e \in \mathcal{E}_T} \int_e \frac{\partial u}{\partial n_e} [v]_e \, ds \quad (137)$$

where \mathcal{E}_T is the set of edges in T that are not on Γ , n_e is a unit vector normal to e , and $[v]_e = v_+ - v_-$ is the jump of v across e (n_e is pointing from the minus side to the plus side and $v = 0$ outside Ω). Since $[v]_e$ vanishes at the midpoint of $e \in \mathcal{E}_T$ we have

$$\int_e \frac{\partial u}{\partial n_e} [v]_e \, ds = \int_e \frac{\partial(u - p)}{\partial n_e} [v]_e \, ds \quad \forall p \in P_1 \quad (138)$$

Let $T_e = \{T \in T : e \subset \partial T\}$. It follows from (138), the trace theorem and the Bramble-Hilbert lemma (cf. Remark 10) that

$$\begin{aligned} \left| \int_e \frac{\partial u}{\partial n_e} [v]_e \, ds \right| &\leq C \inf_{p \in P_1} \left[\sum_{T \in T_e} \left(\|u - p\|_{H^1(T)}^2 + (\text{diam } T)^{2\alpha(T)} |u|_{H^{1+\alpha(T)}(T)}^2 \right) \right]^{1/2} \left(\sum_{T \in T_e} |v|_{H^1(T)}^2 \right)^{1/2} \\ &\leq C \left(\sum_{T \in T_e} (\text{diam } T)^{2\alpha(T)} |u|_{H^{1+\alpha(T)}(T)}^2 \right)^{1/2} \left(\sum_{T \in T_e} |v|_{H^1(T)}^2 \right)^{1/2} \end{aligned} \quad (139)$$

We conclude from (137) and (139) that

$$\sup_{v \in V_T} \frac{a_T(u - u_T, v)}{\|v\|_{a_T}} \leq C \left(\sum_{T \in T} (\text{diam } T)^{2\alpha(T)} |u|_{H^{1+\alpha(T)}(T)}^2 \right)^{1/2} \quad (140)$$

where the constant C depends only on the minimum angle in T .

Combining (135), (136), and (140) we have the following analog of (57)

$$\sum_{T \in T} \|u - u_T\|_{H^1(T)}^2 \leq C \sum_{T \in T} (\text{diam } T)^{2\alpha(T)} |u|_{H^{1+\alpha(T)}(T)}^2 \quad (141)$$

Remark 18. Estimates of $\|u - u_T\|_{L_2(\Omega)}$ can also be obtained for nonconforming finite element methods (Crouzeix and Raviart, 1973). There is also a close connection between certain nonconforming methods and mixed methods (Arnold and Brezzi, 1982).

7.3 Effects of numerical integration

The explicit form of the finite element equation (54) involves the evaluations of integrals which, in general, cannot be computed exactly. Thus, the effects of numerical integration must be taken into account in the error analysis. We will illustrate the ideas in terms of finite element methods for simplicial triangulations. The readers are referred to Davis and Rabinowitz (1984) for a comprehensive treatment of numerical integration.

Consider the second order elliptic boundary value problem (5) in Example 1 on a bounded polyhedral domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3). Let T be a simplicial triangulation of Ω such that Γ is a union of the edges (faces) of T and $V_T \subset V$ be the corresponding P_1 Lagrange finite element space.

In Section 4, the finite element approximate solution $u_T \in V_T$ is defined by (54). But in practice, the integral $F(v) = \int_{\Omega} f v \, dx$ is evaluated by a quadrature scheme and the approximate solution $u_T \in V_T$ is actually defined by

$$a(u_T, v) = F_T(v) \quad \forall v \in V_T \quad (142)$$

where $F_T(v)$ is the result of applying the quadrature scheme to the integral $F(v)$.

More precisely, let $D \in T$ be arbitrary and $\Phi_D : S \rightarrow \bar{D}$ be an affine homeomorphism from the standard (closed) simplex S onto D . It follows from a change of variables that

$$\int_D f v \, dx = \int_S (\det J_{\Phi_D})(f \circ \Phi_D)(v \circ \Phi_D) \, d\bar{x}$$

where without loss of generality $\det J_{\Phi_D}$ (the determinant of the Jacobian matrix of Φ) is assumed to be a positive number. The integral on S is evaluated by a quadrature scheme I_S and the right-hand side of (142) is then given by

$$F_h(v) = \sum_{D \in T} I_S((\det J_{\Phi_D})(f \circ \Phi_D)(v \circ \Phi_D)) \quad (143)$$

The error $u - u_T$ can be estimated by the following analog of (135):

$$\|u - u_T\|_a \leq \inf_{v \in V_T} \|u - v\|_a + \sup_{v \in V_T \setminus \{0\}} \frac{a(u - u_T, v)}{\|v\|_a} \quad (144)$$

The first term on the right-hand side of (144) can be estimated by $\|u - \Pi_T u\|_a$ as in Section 4.1. The second term on the right-hand side of (144) measures the effect of numerical quadrature. Below we give conditions on the quadrature scheme $w \mapsto I_S(w)$ and the function f so that the magnitude of the quadrature error is identical with that of the optimal interpolation error for the finite element space.

We assume that the quadrature scheme $w \mapsto I_S(w)$ has the following properties:

$$|I_S(w)| \leq C_S \max_{\hat{x} \in \hat{S}} |w(\hat{x})| \quad \forall w \in C^0(\hat{S}), \quad (145)$$

$$I_S(w) = \int_{\hat{S}} w \, d\hat{x} \quad \forall w \in P_{2n-2} \quad (146)$$

We also assume that $f \in W_q^n(\Omega)$ such that $q \geq 2$ and $n > d/q$, which implies, in particular, by the Sobolev embedding theorem that $f \in C^0(\Omega)$ so that (143) makes sense.

Under these conditions it can be shown (Ciarlet, 1978) by using the Bramble–Hilbert lemma on S that

$$\left| \int_D f v \, dx - I_S((\det J_{\Phi_D})(f \circ \Phi_D)(v \circ \Phi_D)) \right| \leq C(\text{diam } D)^q |D|^{(1/2)-(1/q)} \|f\|_{W_q^2(D)} \|v\|_{H^1(D)} \quad \forall v \in V_T \quad (147)$$

where the positive constant C depends only on the shape regularity of D . It then follows from (1), (143), (147) and Hölder's inequality that

$$|a(u - u_T, v)| = \left| \int_{\Omega} f v \, dx - F_h(v) \right| \leq C h^q \|f\|_{W_q^2(\Omega)} \|v\|_{H^1(\Omega)} \quad \forall v \in V_T \quad (148)$$

We conclude from (144) and (148) that

$$\|u - u_T\|_a \leq \|u - \Pi_T u\|_a + C h_T^{q+1} \|f\|_{W_q^2(\Omega)} \quad (149)$$

We see by comparing (43) and (149) that the overall accuracy of the finite element method is not affected by the numerical integration.

We now consider a general symmetric positive definite elliptic boundary problem whose variational form is defined

by

$$a(w, v) = \sum_{i,j=1}^d a_{ij}(x) \frac{\partial w}{\partial x_i} \frac{\partial v}{\partial x_j} \, dx + \int_{\Omega} b(x) w v \, dx \quad (150)$$

where $a_{ij}, b \in W_2^1(\Omega)$, $b \geq 0$ on Ω and there exists a positive constant c such that

$$\sum_{i,j=1}^d a_{ij}(x) \xi_i \xi_j \geq c |\xi|^2 \quad \forall x \in \Omega \quad \text{and} \quad \xi \in \mathbb{R}^d \quad (151)$$

In this case, the bilinear form (150) must also be evaluated by numerical integration and the approximation solution $u_T \in V_T$ to (1) is defined by

$$a_T(u_T, v) = F_T(v) \quad \forall v \in V_T \quad (152)$$

where $a_T(w, v)$ is the result of applying the quadrature scheme I_S to the pull-back of

$$\sum_{i,j=1}^d \int_D a_{ij}(x) \frac{\partial w}{\partial x_i} \frac{\partial v}{\partial x_j} \, dx + \int_D b(x) w v \, dx$$

on the standard simplex S under the affine homeomorphism Φ_D , over all $D \in \mathcal{T}$.

The error $u - u_T$ can be estimated under (145), (146) and the additional condition that

$$g \geq 0 \text{ on } \bar{S} \implies I_S(g) \geq 0 \quad (153)$$

Indeed (146), (151), (153) and the sign of b imply that

$$a_T(v, v) \geq c \|v\|_{H^1(\Omega)}^2 \quad \forall v \in V_T \quad (154)$$

and we have the following analog of (144)

$$\begin{aligned} \|u - u_T\|_{H^1(\Omega)} &\leq \inf_{v \in V_T} \|u - v\|_{H^1(\Omega)} \\ &\quad + \frac{1}{c} \left\{ \sup_{v \in V_T \setminus \{0\}} \frac{a_T(u, v) - a(u, v)}{\|v\|_{H^1(\Omega)}} \right. \\ &\quad \left. + \sup_{v \in V_T \setminus \{0\}} \frac{a(u, v) - a_T(u_T, v)}{\|v\|_{H^1(\Omega)}} \right\} \quad (155) \end{aligned}$$

The first term on the right-hand side of (155) is dominated by $\|u - \Pi_T u\|_{H^1(\Omega)}$. Since $a(u, v) - a_T(u_T, v) = \int_{\Omega} f v \, dx - F_T(v)$, the third term is controlled by the estimate (148). The second term, which measures the effect of

numerical quadrature on the bilinear form $a(\cdot, \cdot)$, is controlled by the estimate

$$|a_T(u, v) - a(u, v)| \leq C \sum_{D \in \mathcal{T}} (\text{diam } D)^{\alpha(D)} \times |u|_{H^{1+\alpha(D)}(D)} \|v\|_{H^1(D)} \quad (156)$$

provided the solution u belongs to $H^{1+\alpha(D)}(D)$ for each $D \in \mathcal{T}$ and $1/2 < \alpha(D) \leq 1$. The estimate (156) follows from (145), (146) and the Bramble–Hilbert lemma, and the positive constant C in (156) depends only on the W_2^2 norms of a_{ij} and b and the shape regularity of \mathcal{T} . Again we see by comparing (56), (149), and (156) that the overall accuracy of the finite element method is not affected by the numerical integration.

Remark 19. For problems that exhibit the phenomenon of locking, the choice of a lower order quadrature scheme in the evaluation of the stiffness matrix may help alleviate the effect of locking (Malkus and Hughes, 1978).

7.4 Curved domains

So far, we have restricted the discussion to polygonal (polyhedral) domains. In this section, we consider the second-order elliptic boundary value problem (5) on a domain $\Omega \subset \mathbb{R}^2$ with a curved boundary. For simplicity, we assume that $\Gamma = \partial\Omega$.

First, we consider the P_1 Lagrange finite element method for a domain Ω with a C^2 boundary. We approximate Ω by a polygonal domain Ω_h on which a simplicial triangulation \mathcal{T}_h of mesh size h is imposed. We assume that the vertices of \mathcal{T}_h belong to the closure of Ω . A typical triangle in \mathcal{T}_h near $\partial\Omega$ is depicted in Figure 20(a).

Let $V_h \subset H_0^1(\Omega_h)$ be the P_1 finite element space associated with \mathcal{T}_h . The approximate solution $u_h \in V_h$ for (5) is then defined by

$$a_h(u_h, v) = F_h(v) \quad \forall v \in V_h \quad (157)$$

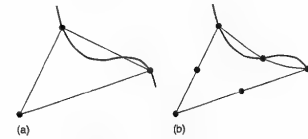


Figure 20. Triangulations for curved domains.

where

$$a_h(w, v) = \int_{\Omega_h} \nabla w \cdot \nabla v \, dx \quad \forall w, v \in H^1(\Omega_h) \quad (158)$$

and $F_h(v)$ represented the result of applying a numerical quadrature scheme to the integral $\int_{\Omega_h} f v \, dx$ (cf. (143) with f and T replaced by \tilde{f} and T_h). Here \tilde{f} is an extension of f to \mathbb{R}^2 . We assume that the numerical scheme uses only the values of \tilde{f} at the nodes of T_h and hence the discrete problem (157) is independent of the extension \tilde{f} .

We assume that $f \in W_q^2(\Omega)$ for $2 < q < \infty$ and hence $u \in W_q^2(\Omega)$ by elliptic regularity. Let $\tilde{u} \in W_q^2(\mathbb{R}^2)$ be an extension of u such that

$$\|\tilde{u}\|_{W_q^2(\mathbb{R}^2)} \leq C_{\Omega} \|u\|_{W_q^2(\Omega)} \leq C_{\Omega} \|f\|_{W_q^2(\Omega)} \quad (159)$$

We can then take $\tilde{f} = -\Delta \tilde{u} \in W_q^1(\mathbb{R}^2)$ to be the extension appearing in the definition of F_h .

The error $\tilde{u} - u_h$ over Ω_h can be estimated by the following analog of (135):

$$\|\tilde{u} - u_h\|_{a_h} \leq \inf_{v \in V_h} \|\tilde{u} - v\|_{a_h} + \sup_{v \in V_h \setminus \{0\}} \frac{a_h(\tilde{u} - u_h, v)}{\|v\|_{a_h}} \quad (160)$$

The first term on the right-hand side is dominated by $\|\tilde{u} - \Pi_h \tilde{u}\|_{a_h}$ where Π_h is the nodal interpolation operator. The second term is controlled by

$$|a_h(\tilde{u} - u_h, v)| = \left| \int_{\Omega_h} \tilde{f} v \, dx - F_h(v) \right| \leq Ch \|\tilde{f}\|_{W_q^1(\Omega_h)} \|v\|_{H^1(\Omega_h)} \quad (161)$$

which is a special case of (148), provided that the conditions (145) and (146) (with $n = 1$) on the numerical quadrature scheme are satisfied.

Combining (43) and (159)–(161) we see that

$$\|\tilde{u} - u_h\|_{H^1(\Omega_h)} = \|\tilde{u} - u_h\|_{a_h} \leq Ch \|f\|_{W_q^2(\Omega)} \quad (162)$$

that is, the P_1 finite element method for the curved domain retains the optimal $O(h)$ accuracy.

The approximation of Ω_h to Ω can be improved if we replace straight-edge triangles by triangles with a curved edge (Figure 20(b)). This can be achieved by the *isoparametric* finite element methods. We will illustrate the idea using the P_2 Lagrange element.

Let Ω_h an approximation of Ω , be the union of straight-edge triangles (in the interior of Ω_h) and triangles with one curved edge (at the boundary of Ω_h), which form a triangulation \mathcal{T}_h of Ω_h . The finite element associated with the interior triangles is the standard P_2 Lagrange element.

For a triangle D at the boundary, we assume that there is a homeomorphism Φ_D from the standard simplex S onto D such that $\Phi_D(\hat{x}) = (\Phi_{D,1}(\hat{x}), \Phi_{D,2}(\hat{x}))$ where $\Phi_{D,1}(\hat{x})$ and $\Phi_{D,2}(\hat{x})$ are quadratic polynomials in $\hat{x} = (\hat{x}_1, \hat{x}_2)$. The space of shape functions \mathcal{P}_D is then defined by

$$\mathcal{P}_D = \{v \in C^\infty(D) : v \circ \Phi_D \in \mathcal{P}_2(S)\} \quad (163)$$

that is, the functions in \mathcal{P}_D are quadratic polynomials in the curvilinear coordinates on D induced by Φ_D^{-1} . The set \mathcal{N}_D of nodal variables consist of pointwise evaluations of the shape functions at the nodes corresponding to the nodes of the P_2 element on S (cf. Figures 2 and 20) under the map Φ_D . We assume that all such nodes belong to $\bar{\Omega}$ and the nodes on the curved edge of D belong to $\partial\Omega$ (cf. Figure 20).

In other words, the finite element $(D, \mathcal{P}_D, \mathcal{N}_D)$ is pulled back to the P_2 Lagrange finite element on S under Φ_D . It is called an isoparametric element because the components of the parameterization map Φ_D belong to the shape functions of the P_2 element on S . The corresponding finite element space defined by (32) (with T replaced by T_h) is a subspace of $H^1(\Omega_h)$ that contains all the continuous piecewise linear polynomials with respect to T_h . By setting the nodal values on $\partial\Omega_h$ to be zero, we have a finite element space $V_h \subset H_0^1(\Omega_h)$. The discrete problem for $u_h \in V_h$ is then defined by

$$\tilde{a}_h(u_h, v) = F_h(v) \quad (164)$$

where the numerical quadrature scheme in the definition of F_h involves only the nodes of the finite element space so that the discrete problem is independent of the choice of the extension of f and the variational form $\tilde{a}_h(\cdot, \cdot)$ is obtained from $a_h(\cdot, \cdot)$ by the numerical quadrature scheme. *

We assume that $f \in W_q^2(\Omega)$ for $1 < q < \infty$ and hence $u \in W_q^2(\Omega)$ by elliptic regularity (assuming that Ω has a C^3 boundary). Let $\tilde{u} \in W_q^2(\mathbb{R}^2)$ be an extension of u such that

$$\|\tilde{u}\|_{W_q^2(\mathbb{R}^2)} \leq C_\Omega \|u\|_{W_q^2(\Omega)} \leq C_\Omega \|f\|_{W_q^2(\Omega)} \quad (165)$$

Under the condition (153), we have

$$\tilde{a}_h(v, v) \geq c \|v\|_{H^1(\Omega_h)}^2 \quad \forall v \in V_h \quad (166)$$

and the error of $\tilde{u} - u_h \in \Omega_h$ can be estimated by the following analog of (155)

$$\begin{aligned} \|\tilde{u} - u_h\|_{H^1(\Omega_h)} &\leq \inf_{v \in V_h} \|\tilde{u} - v\|_{H^1(\Omega_h)} \\ &+ \frac{1}{c} \left\{ \sup_{v \in V_h \setminus \{0\}} \frac{\tilde{a}_h(\tilde{u}, v) - a_h(\tilde{u}, v)}{\|v\|_{H^1(\Omega_h)}} \right\} \end{aligned}$$

$$+ \sup_{v \in V_h \setminus \{0\}} \frac{a_h(\tilde{u}, v) - \tilde{a}_h(u_h, v)}{\|v\|_{H^1(\Omega_h)}} \quad (167)$$

The analysis of the terms on the right-hand side of (167) involves the shape regularity of a curved triangle, which can be defined as follows. Let \mathcal{A}_D be the affine map that agrees with Φ_D at the vertices of the standard simplex S , and \tilde{D} be the image of S under \mathcal{A}_D . (\tilde{D} is the triangle in Figure 20(a), while D is the curved triangle in (b).) The shape regularity of the curved triangle D is measured by the aspect ratio $\gamma(\tilde{D})$ (cf. (29)) of the straight-edged triangle \tilde{D} and the parameter $\kappa(D)$ defined by

$$\kappa(D) = \max \{h^{-1} |\Phi_D \circ \mathcal{A}_D^{-1}|_{W_\infty^2(D)}, h^{-2} |\Phi_D \circ \mathcal{A}_D^{-1}|_{W_\infty^3(S)}\} \quad (168)$$

and we can take the aspect ratio $\gamma(D)$ to be the maximum of $\gamma(\tilde{D})$ and $\kappa(D)$. Note that in the case where $D = \tilde{D}$ the parameter $\kappa(D) = 0$ and $\gamma(D) = \gamma(\tilde{D})$.

The first term on the right-hand side of (167) is dominated by $\|\tilde{u} - \Pi_h \tilde{u}\|_{\Omega_h}$, where Π_h is the nodal interpolation operator. Note that, by using the Bramble-Hilbert lemma on S and scaling, we have the following generalization of (43):

$$\|\tilde{u} - \Pi_D \tilde{u}\|_{H^1(D)} \leq C(\text{diam } D)^2 \|\tilde{u}\|_{H^3(D)} \quad (169)$$

where Π_D is the element nodal interpolation operator and the constant C depends only on an upper bound of $\gamma(D)$, and hence

$$\|\tilde{u} - \Pi_h \tilde{u}\|_{H^1(\Omega_h)} \leq Ch^2 \|\tilde{u}\|_{H^3(\Omega_h)} \quad (170)$$

In order to analyze the third term on the right-hand side of (167), we take $\tilde{f} = -\Delta \tilde{u}$ and impose the conditions (145) and (146) (with $n = 2$) on the numerical quadrature scheme. We then have the following special case of (148):

$$\begin{aligned} |a_h(\tilde{u}, v) - \tilde{a}_h(u_h, v)| &= \left| \int_{\Omega_h} \tilde{f} v \, dx - F_h(v) \right| \\ &\leq Ch^2 \|\tilde{f}\|_{W_q^1(\Omega_h)} \|v\|_{H^1(\Omega_h)} \end{aligned} \quad (171)$$

Similarly, the second term on the right-hand side of (167), which measures the effect of numerical integration on the variational form $a_h(\cdot, \cdot)$, is controlled by the estimate

$$|\tilde{a}_h(\tilde{u}, v) - a_h(\tilde{u}, v)| \leq Ch^2 \|\tilde{u}\|_{H^2(\Omega_h)} \|v\|_{H^1(\Omega_h)} \quad \forall v \in V_h \quad (172)$$

Combining (167) and (170)–(172) we have

$$\|\tilde{u} - u_h\| \leq Ch^2 \|f\|_{W_q^2(\Omega)} \quad (173)$$

where C depends only on an upper bound of $\{\gamma(D) : D \in \mathcal{T}_h\}$ and the constants in (165). Therefore, the P_2 isoparametric finite element method retains the optimal $O(h^2)$ accuracy. On the other hand, if only straight-edged triangles are used in the construction of Ω_h , then the accuracy of the P_2 Lagrange finite element method is only of order $O(h^{3/2})$ (Strang and Berger, 1971).

The discussion above can be generalized to higher-order isoparametric finite element methods, higher dimensions, and elliptic problems with variable coefficients (Ciarlet, 1978).

Remark 20. Estimates such as (160) and (162) are useful only when a sequence of domains Ω_h with corresponding triangulations T_h can be constructed so that $h_i \downarrow 0$ and the aspect ratios of all the triangles (straight or curved) in the triangulations remain bounded. We refer the readers to Scott (1973) for 2-D constructions and to Lenoir (1986) for the 3-D case.

Remark 21. Other finite element methods for curved domains can be found in Zlámal (1973, 1974), Scott (1975), and Bernardi (1989).

Remark 22. Let Ω_i be a sequence of convex polygons approaching the unit disc. The displacement of the simply supported plate on Ω_i with unit loading does not converge to the displacement of the simply supported plate on the unit disc (also with unit loading) as $i \rightarrow \infty$. This is known as Babuška's plate paradox (Babuška and Pitkäranta, 1990). It shows that numerical solutions obtained by approximating a curved domain with polygonal domains, in general, do not converge to the solution of a fourth-order problem defined on the curved domain. We refer the readers to Mansfield (1978) for the construction of finite element spaces that are subspaces of $H^2(\Omega)$.

7.5 Pointwise estimates

Besides the estimates in L_2 -based Sobolev spaces discussed in Section 4, there also exist a priori error estimates for finite element methods in L_p -based Sobolev spaces with $p \neq 2$. In particular, error estimates in the L_∞ -based Sobolev spaces can provide pointwise error estimates. Below we describe some results for second-order elliptic boundary value problems with homogeneous Dirichlet boundary conditions.

In the one-dimensional case (Wheeler, 1973) where Ω is an interval, the finite element solution u_T for a given

triangulation T with mesh size h_T satisfies

$$\|u - u_T\|_{L_\infty(\Omega)} \leq Ch_T^2 \|u\|_{W_\infty^3(\Omega)} \quad (174)$$

provided the solution u of (5) belongs to $W_\infty^3(\Omega)$ and the finite element space contains all the piecewise polynomial functions of degree $\leq n - 1$. The estimate (174) also holds in higher dimensions (Douglas, Dupont and Wheeler, 1974) in the case where Ω is a product of intervals and u_T is the solution in the Q_{n-1} finite element space of Example 7.

For a two-dimensional convex polygonal domain (Natterer, 1975; Scott, 1976; Nitsche, 1977), the estimate (174) holds in the case where $n \geq 3$ and u_T is the P_{n-1} triangular finite element solution for a general triangulation T . In the case where u_T is the P_1 finite element solution, (174) is replaced by

$$\|u - u_T\|_{L_\infty(\Omega)} \leq Ch_T^2 |\ln h_T| \|u\|_{W_\infty^3(\Omega)} \quad (175)$$

L_∞ estimates for general triangulations on polygonal domains with reentrant corners and higher-dimensional domains can be found in Schatz and Wahlbin (1978, 1979, 1982) and Schatz (1998). The estimate (175) was also established in Gastaldi and Nochetto (1987) for the Crouzeix-Raviart nonconforming P_1 element of Example 16.

It is also known (Rannacher and Scott, 1982; Brenner and Scott, 2002) that

$$\|u - u_T\|_{W_\infty^1(\Omega)} \leq C \inf_{v \in V_T} \|u - v\|_{W_\infty^1(\Omega)} \quad (176)$$

where Ω is a convex polygonal domain in \mathbb{R}^2 and u_T is the P_n ($n \geq 1$) triangular finite element solution obtained from a general triangulation T of Ω . Optimal order estimates for $\|u - u_T\|_{W_\infty^1(\Omega)}$ can be derived immediately from (176). Extension of (176) to higher dimensions can be found in Schatz and Wahlbin (1995).

7.6 Interior estimates and pollution effects

Let Ω be the L -shaped polygon in Figure 1. The solution u of the Poisson problem (5) on Ω with homogeneous Dirichlet boundary condition is singular near the reentrant corner and $u \notin H^2(\Omega)$. Consequently, the error estimate $\|u - u_T\|_{H^1(\Omega)} \leq Ch_T \|f\|_{L_2(\Omega)}$ does not hold for the P_1 triangular finite element solution u_T associated with a quasi-uniform triangulation T of mesh size h_T .

However, u does belong to $H^2(\Omega_\delta)$ where Ω_δ is the subset of the points of Ω whose distances to the reentrant

corner are strictly greater than the positive number δ . Therefore, it is possible that

$$\|u - u_T\|_{H^1(\Omega)} \leq Ch_T \|f\|_{L^2(\Omega)} \quad (177)$$

That the estimate (177) indeed holds is a consequence of the following interior estimate (Nitsche and Schatz, 1974):

$$\|u - u_T\|_{H^1(\Omega)} \leq C \left(\inf_{v \in V_T} \|u - v\|_{H^1(\Omega_h)} + \|u - u_T\|_{L^2(\Omega_h)} \right) \quad (178)$$

where $V_T \subset H_0^1(\Omega)$ is the P_1 triangular finite element space. Interior estimates in various Sobolev norms can be established for subdomains of general Ω in \mathbb{R}^d and general finite elements. We refer the readers to Wahlbin (1991) for a survey of such results and to Schatz (2000) for some recent developments.

On the other hand, since u_T is obtained by solving a global system that involves the nodal values near the reentrant corner of the L -shaped domain, the effect of the singularity at the reentrant corner can propagate into other parts of Ω . This is known as the *pollution effect* and is reflected, for example, by the following estimate (Wahlbin,

1984):

$$\|u - u_T\|_{L^2(\Omega)} \geq Ch_T^{\frac{2\beta}{3}} \quad (179)$$

where $\beta = \pi/(3\pi/2) = 2/3$. Similar estimates can also be established for other Sobolev norms.

7.7 Superconvergence

Let u_T be the finite element solution of a second-order elliptic boundary value problem. Suppose that the space of shape functions on each element contains all the polynomials of degree $\leq n$ but not all the polynomials of degree $n+1$. Then the L_∞ norm of the error $u - u_T$ is at most of order h^{n+1} , even if the solution u is smooth. However, the absolute value of $u - u_T$ at certain points can be of order $h^{n+1+\sigma}$ for some $\sigma > 0$. This is known as the phenomenon of *superconvergence* and such points are the superconvergence points for u_T . Similarly, a point where the absolute value of a derivative of $u - u_T$ is of order $h^{n+1+\sigma}$ is a superconvergence point for the derivative of u_T .

The division points of a partition T for a two point boundary value problem with smooth coefficients provides

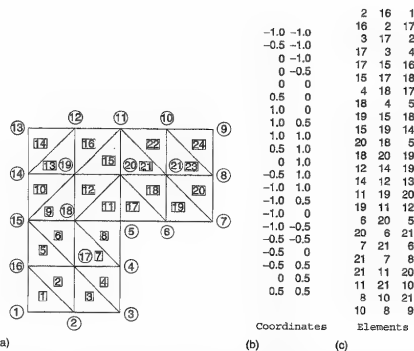


Figure 21. Picture of a triangulation $T = \text{conv}\{(-1/2, -1), (-1, -1/2), (-1, -1), \text{conv}\{(-1, -1/2), (-1/2, -1), (-1/2, -1/2)\}, \dots, \text{conv}\{(1/2, 1), (1, 1/2), (1, 1)\}$ with $m = 24$ triangles and $n = 21$ nodes (a). The picture indicates an enumeration of nodes (numbers in circles) and elements (numbers in boxes) given in the matrices coordinates (b) and elements (c). The Dirichlet boundary conditions on the exterior nodes are included in the vector $\text{dirichlet} = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16)$ of the labels in a counterclockwise enumeration. The data coordinates , elements , and dirichlet are the input of the finite element program to compute a displacement vector x as its output.

the simplest example of superconvergence points. Let u_T be the finite element solution from the P_n Lagrange finite element space. Since the Green's function G_p associated with a division point p is continuous in the interval Ω and smooth on the two subintervals divided by p , we have (Douglas and Dupont, 1974)

$$\begin{aligned} |(u - u_T)(p)| &= |a(u - u_T, G_p)| \\ &= |a(u - u_T, G_p - \Pi_T^n G_p)| \\ &\leq C \|u - u_T\|_{H^1(\Omega)} \|G_p - \Pi_T^n G_p\|_{H^1(\Omega)} \leq Ch^{2n} \end{aligned}$$

provided that u is sufficiently smooth. Therefore, p is a superconvergence point for u_T if $n \geq 2$.

For general superconvergence results in various dimensions, we refer the readers to Krížek and Neittaanmäki (1987), Chen and Huang (1995), Wahlbin (1995), Lin and Yan (1996), Schatz, Sloan and Wahlbin (1996), Krížek, Neittaanmäki and Stenberg (1998), Chen (1999), and Babuška and Strouboulis (2001).

7.8 Finite element program in 15 lines of MATLAB

It is the purpose of this section to introduce a short (two-dimensional) P_1 finite element program.

The data for a given triangulation $T = \{T_1, \dots, T_m\}$ into triangles with a set of nodes $N = \{z_1, \dots, z_n\}$ are described in user-specified matrices called *coordinates* and *elements*. Figure 21 displays a triangulation with m triangles and n nodes as well as a fixed enumeration and the corresponding data. The coordinates of the nodes $z_k = (x_k, y_k)$ (d real components in general) are stored in the k th row of the two-dimensional matrix *coordinates*. Each element $T_j = \text{conv}\{z_k, z_\ell, z_m\}$ is represented by the labels of its vertices (k, ℓ, m) stored in the j th row of the two-dimensional matrix *elements*. The chosen permutation of (k, ℓ, m) describes the element in a counterclockwise orientation. Homogeneous Dirichlet conditions are prescribed on the boundary specified by an input vector *dirichlet* of all fixed nodes at the outer boundary; cf. Figure 21.

Given the aforementioned data in the model Dirichlet problem with right-hand side $f = 1$, the P_1 finite element space $V := \text{span}\{\phi_j : z_j \in K\}$ is formed by the nodal basis functions ϕ_j of each free node z_k ; the set K of free nodes, the interior nodes, is represented in the N vector *freedomes*, the vector of labels in $1:n$ without *dirichlet*.

The resulting discrete equation is the $N \times N$ linear system of equations $Ax = b$ with the positive definite

symmetric stiffness matrix A and right-hand side b . Their components are defined (as a subset of)

$$A_{jk} := \int_{\Omega} \nabla \phi_j \cdot \nabla \phi_k \, dx$$

and

$$b_j := \int_{\Omega} f \phi_j \, dx \quad \text{for } j, k = 1, \dots, n$$

The computation of the entries A_{jk} and b_j is performed elementwise for the additivity of the integral and since T is a partition of the domain Ω . Given the triangle T_j number j , the MATLAB command *coordinates*(*elements*(*j*, :), :) returns the 3×2 matrix $(P_1, P_2, P_3)^T$ of its vertices. Then, the *local stiffness matrix* reads

$$\text{STIMA}(T_j)_{\alpha\beta} := \int_{T_j} \nabla \phi_k \cdot \nabla \phi_\ell \, dx \quad \text{for } \alpha, \beta = 1, 2, 3$$

for those numbers k and ℓ of two vertices $z_k = P_\alpha$ and $z_\ell = P_\beta$ of T_j . The correspondence of global and local indices, i.e. the numbers of vertices in $(z_k, z_\ell, z_m) = (P_1, P_2, P_3)$, of T_j can be formalized by

$$I(T_j) = \{(\alpha, k) \in \{1, 2, 3\} \times \{1, \dots, n\} : P_\alpha = z_k \in \mathcal{N}\}$$

The local stiffness matrix is in fact

$$\text{STIMA}(T_j) = \det \frac{P}{2} (Q Q^T) \text{ with } P := \begin{bmatrix} 1 & 1 & 1 \\ P_1 & P_2 & P_3 \end{bmatrix}$$

$$\text{and } Q := P^{-1} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This formula allows a compact programming in MATLAB as shown (for any dimension d)

```
function stima=stima(vertices)
P=[ones(1,size(vertices,2)+1);vertices'];
Q=2*[zeros(1,size(vertices,2))...
eye(size(vertices,2))];
stima=det(P)*Q*Q'/prod(1:size(vertices,2));
```

Utilizing the index sets I , the assembling of all local stiffness matrices reads

$$\text{STIMA} = \sum_{T_j \in \mathcal{T}} \sum_{(\alpha,k) \in I(T_j)} \sum_{(\beta,\ell) \in I(T_j)} \text{STIMA}(T_j)_{\alpha\beta} e_k \otimes e_\ell$$

(e_k is the k th canonical unit vector with the ℓ th component equal to the Kronecker delta $\delta_{k\ell}$ and \otimes is the dyadic product.) The implementation of each summation is realized by adding $\text{STIMA}(T_j)$ to the 3×3 submatrix of the rows and columns corresponding to k, ℓ, m ; see the MATLAB program below.

```

function x=FEM(coordinates,elements,dirichlet)
A=sparse(size(coordinates,1),size(coordinates,1));
b=sparse(size(coordinates,1),1);x=zeros(size(coordinates,1),1);
for j=1:size(elements,1)
    A(elements(j,:),elements(j,:))=A(elements(j,:),elements(j,:))+
    +stiffness(coordinates(elements(j,:),:));
    b(elements(j,:))=b(elements(j,:))+ones(3,1)...
    +dist([1,1,1;coordinates(elements(j,:),:)]')/6;
end
freeNodes=setdiff(1:size(coordinates,1),dirichlet);
x(freeNodes)=A(freeNodes,freeNodes)\b(freeNodes);

```

Given the output vector x , a plot of the discrete solution

$$\tilde{u} = \sum_{j=1}^n x_j \varphi_n$$

is generated by the command `trisurf(elements,coordinates(:,1),coordinates(:,2),x)` and displayed in Figure 22.

For alternative programs with numerical examples and full documentation, the interested readers are referred to Albrecht, Carstensen and Funken (1999) and Albrecht *et al.* (2002). The closest more commercial finite element package might be FEMLAB. The internet provides over 200 000 entries under the search for 'Finite Element Method Program'. Amongst public domain software are the programs ALBERT (Freiburg), DEAL (Heidelberg), UG (Heidelberg), and so on.

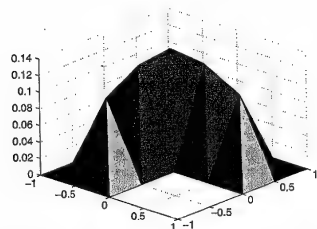


Figure 22. Discrete solution of $-\Delta u = 1$ with homogeneous Dirichlet boundary data based on the triangulation of Figure 21. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

ACKNOWLEDGMENTS

The work of Susanne C. Brenner was partially supported by the National Science Foundation under Grant Numbers DMS-00-74246 and DMS-03-11790. The work of Carsten Carstensen has been initiated while he was visiting the Isaac-Newton Institute of Mathematical Sciences, Cambridge, England; the support of the EPSRC under grant N09176/01 is thankfully acknowledged. The authors thank Dipl.-Math. Jan Bolte for his assistance with the numerical examples of this chapter.

REFERENCES

- Adams RA. *Sobolev Spaces*. Academic Press, New York, 1995.
- Agmon S. *Lectures on Elliptic Boundary Value Problems*. Van Nostrand, Princeton, 1965.
- Ainsworth M and Oden JT. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley-Interscience, New York, 2000.
- Albrecht J, Carstensen C and Funken S. Remarks around 50 lines of Matlab: Short finite element implementation. *Numer. Algorithms* 1999; 20:117–137.
- Albrecht J, Carstensen C, Funken S and Klose R. Matlab implementation of the finite element method in elasticity. *Computing* 2002; 60:239–263.
- Apel T. *Anisotropic Finite Elements: Local Estimates and Applications*. Teubner Verlag, Stuttgart, 1999.
- Apel T and Dobrowolski M. Anisotropic interpolation with applications to the finite element method. *Computing* 1992; 47:277–293.
- Arnold DN and Brezzi F. Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO Modél. Math. Anal. Numér.* 1982; 19:7–32.
- Arnold DN, Boffi D and Falk RS. Approximation by quadrilateral finite elements. *Math. Comput.* 2002; 71:909–922.
- Arnold DN, Mukherjee A and Pouly L. Locally adapted tetrahedral meshes using bisection. *SIAM J. Sci. Comput.* 2000; 22:431–448.
- Aziz AK (ed.). *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*. Academic Press, New York, 1972.
- Babuška I. Courant element: before and after. *Lecture Notes in Pure and Applied Mathematics*, vol. 164. Marcel Dekker: New York, 1994; 37–51.
- Babuška I and Aziz AK. Survey lectures on the mathematical foundations of the finite element method. In *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Aziz AK (ed.). Academic Press, New York, 1972; 3–359.
- Babuška I and Aziz AK. On the angle condition in the finite element method. *SIAM J. Numer. Anal.* 1976; 13:214–226.
- Babuška I and Guo B. The h , p , and $h-p$ versions of the finite element methods in 1 dimension. *Numer. Math.* 1986; 49:613–657.
- Babuška I and Kellogg RB. Nonuniform error estimates for the finite element method. *SIAM J. Numer. Anal.* 1975; 12:868–875.
- Babuška I and Miller A. A feedback finite element method with a posteriori error estimation. I. The finite element method and some properties of the a posteriori estimator. *Comput. Methods Appl. Mech. Eng.* 1987; 61:1–40.
- Babuška I and Osborn J. Eigenvalue problems. In *Handbook of Numerical Analysis*, vol. II, Claret PG and Lions JL (eds). North Holland: Amsterdam, 1991; 641–787.
- Babuška I and Pitkäranta J. The plate paradox for hard and soft simple support. *SIAM J. Math. Anal.* 1990; 21:551–576.
- Babuška I and Rheinboldt WC. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.* 1978; 15:736–754.
- Babuška I and Rheinboldt WC. Analysis of optimal finite-element meshes in \mathbb{R}^1 . *Math. Comput.* 1979; 33:435–463.
- Babuška I and Strohobius T. *The Finite Element Method and its Reliability*. Oxford University Press, New York, 2001.
- Babuška I and Suri M. On locking and robustness in the finite element method. *SIAM J. Numer. Anal.* 1992; 29:1261–1293.
- Babuška I and Vogelius R. Feedback and adaptive finite element solution of one-dimensional boundary value problems. *Numer. Math.* 1984; 44:75–102.
- Bangerth W and Rannacher R. *Adaptive Finite Element Methods for Differential Equations*, Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2003.
- Bank RE and Weiser A. Some a posteriori error estimators for elliptic differential equations. *Math. Comput.* 1985; 44:283–301.
- Bank RE and Xu J. Asymptotically Exact a Posteriori Error Estimators, Part I: Grids with Superconvergence. *SIAM J. Numer. Anal.* 2003; 41:2294–2312.
- Bänsch E. Local mesh refinement in 2 and 3 dimensions. *IMPACT Comput. Sci. Eng.* 1991; 3:181–191.
- Bartels S and Carstensen C. Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. II. Higher order FEM. *Math. Comput.* 2002; 71:971–994.
- Bathe K-J. *Finite Element Procedures*. Prentice Hall, Upper Saddle River, 1996.
- Becker EB, Carey GF and Oden JT. *Finite Elements. An Introduction*. Prentice Hall, Englewood Cliffs, 1981.
- Becker R and Rannacher R. A feedback approach to error control in finite element methods: basic analysis and examples. *East-West J. Numer. Math.* 1996; 4:237–264.
- Becker R and Rannacher R. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numer.* 2001; 10:1–102.
- Ben Belgacem F and Brenner SC. Some nonstandard finite element estimates with applications to 3D Poisson and Signorini problems. *Electron. Trans. Numer. Anal.* 2001; 12:134–148.
- Berger A, Scott R and Strang G. Approximate boundary conditions in the finite element method. *Symposia Mathematica*, vol. X (Convegno di Analisi Numerica, INDAM, Rome, 1972). Academic Press: London, 1972; 295–313.
- Bernardi C. Optimal finite-element interpolation on curved domains. *SIAM J. Numer. Anal.* 1989; 26:1212–1240.
- Bernardi C and Girault V. A local regularization operator for triangular and quadrilateral finite elements. *SIAM J. Numer. Anal.* 1998; 35:1893–1916.
- Bey J. Tetragonal grid refinement. *Computing* 1995; 55:355–378.
- Binev P, Dahmen W and DeVore R. Adaptive finite element methods with convergence rates. *Numer. Math.* 2004; 97:219–268.
- Braess D. *Finite Elements* (2nd edn). Cambridge University Press, Cambridge, 2001.
- Bramble JH and Hilbert AH. Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolation. *SIAM J. Numer. Anal.* 1970; 7:113–124.
- Bramble JH, Pasciak JE and Schatz AH. The construction of preconditioners for elliptic problems by substructuring. I. *Math. Comput.* 1996; 47:103–134.
- Bramble JH, Pasciak JE and Steinbach O. On the stability of the L_2 projection in $H^1(\Omega)$. *Math. Comput.* 2002; 71:147–156.
- Brenner SC and Scott LR. *The Mathematical Theory of Finite Element Methods* (2nd edn). Springer-Verlag, New York, 2002.
- Brenner SC and Sung LY. Discrete Sobolev and Poincaré inequalities via Fourier series. *East-West J. Numer. Math.* 2000; 8:83–92.
- Carstensen C. Quasi-interpolation and a posteriori error analysis in finite element methods. *M2AN Math. Model. Numer. Anal.* 1999; 33:1187–1202.
- Carstensen C. Merging the Bramble-Pasciak-Steinbach and the Crouzeix-Thomée criterion for H^1 -stability of the L_2 -projection onto finite element spaces. *Math. Comput.* 2002; 71:157–163.
- Carstensen C. All first-order averaging techniques for a posteriori finite element error control on unstructured grids are efficient and reliable. *Math. Comput.* 2004; 73:1153–1165.
- Carstensen C. An adaptive mesh-refining algorithm allowing for an H^1 -stable L_2 -projection onto Courant finite element spaces. *Constructive Approximation Theory*. Newton Institute, 2003b; DOI: 10.1007/s00365-003-0550-5. Preprint N03004-CPD available at <http://www.newton.cam.ac.uk/preprints/N03004.pdf>.
- Carstensen C. Some remarks on the history and future of averaging techniques in a posteriori finite element error analysis. *ZAMM* 2004; 84:3–21.
- Carstensen C and Albrecht J. Averaging techniques for reliable a posteriori FE-error control in elastoplasticity with hardening. *Comput. Methods Appl. Mech. Eng.* 2003; 192:1435–1450.

Carstensen C and Bartels S. Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. I. Low order conforming, nonconforming and Mixed FEM. *Math. Comput.* 2002; 71:945–969.

Carstensen C and Funken SA. Fully reliable localised error control in the FEM. *SIAM J. Sci. Comput.* 1999/00; 21:1465–1484.

Carstensen C and Funken SA. Constants in Clément-interpolation error and residual based a posteriori estimates in finite element methods. *East-West J. Numer. Math.* 2000; 8:153–175.

Carstensen C and Funken SA. A posteriori error control in low-order finite element discretizations of incompressible stationary flow problems. *Math. Comput.* 2001a; 70:1353–1381.

Carstensen C and Funken SA. Averaging technique for FE-a posteriori error control in elasticity. I. Conforming FEM. II. λ -independent estimates. III. Locking-free nonconforming FEM. *Comput. Methods Appl. Mech. Eng.* 2001b; 190:2483–2498; 190:4663–4675; 191:861–877.

Carstensen C and Verfürth R. Edge residuals dominate a posteriori error estimates for low order finite element methods. *SIAM J. Numer. Anal.* 1999; 36:1571–1587.

Carstensen C, Bartels S and Klose R. An experimental survey of a posteriori Courant finite element error control for the Poisson equation. *Adv. Comput. Math.* 2001; 15:79–106.

Chen C. Superconvergence for triangular finite elements. *Sci. China (Ser. A)* 1999; 42:917–924.

Chen CM and Huang YQ. *High Accuracy Theory of Finite Element Methods*. Hunan Scientific and Technical Publisher, Changsha, 1995 (in Chinese).

Ciarlet PG. *The Finite Element Method for Elliptic Problems*. North Holland, Amsterdam, 1978 (Reprinted in the Classics in Applied Mathematics Series, SIAM, Philadelphia, 2002).

Ciarlet PG. *Mathematical Elasticity, Volume I: Three-dimensional Elasticity*. North Holland, Amsterdam, 1988.

Ciarlet PG. Basic error estimates for elliptic problems. In *Handbook of Numerical Analysis*, vol. II, Ciarlet PG and Lions JL (eds). North Holland: Amsterdam, 1991; 17–351.

Ciarlet PG. *Mathematical Elasticity, Volume II: Theory of Plates*. North Holland, Amsterdam, 1997.

Clément P. Approximation by finite element functions using local regularization. *RAIRO Modél. Math. Anal. Numér.* 1975; 9:77–84.

Courant R. Variational methods for the solution of problems of equilibrium and vibration. *Bull. Am. Math. Soc.* 1943; 49:1–23.

Crouzeix M and Raviart PA. Conforming and nonconforming finite element methods for solving the stationary Stokes equations I. *RAIRO Modél. Math. Anal. Numér.* 1975; 7:33–75.

Crouzeix M and Thomée V. The stability in L^p and $W^{1,p}$ of the L^2 -projection onto finite element function spaces. *Math. Comput.* 1987; 48:521–532.

Dauge M. *Elliptic Boundary Value Problems on Corner Domains*. Lecture Notes in Mathematics 1341. Springer-Verlag, Berlin, 1988.

Davis PJ and Rabinowitz P. *Methods of Numerical Integration*. Academic Press, Orlando, 1984.

Demkowicz L. *hp*-adaptive finite elements for time-harmonic Maxwell equations. *Topics in Computational Wave Propagation*, Lecture Notes in Computational Science and Engineering,

pp. 163–199, Ainsworth M, Davies P, Duncan D, Martin P and Rynne B (eds). Springer-Verlag, Berlin, 2003.

Dörfler W. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.* 1996; 33:1106–1124.

Dörfler W and Nochetto RH. Small data oscillation implies the saturation assumption. *Numer. Math.* 2002; 91:1–12.

Douglas Jr J and Dupont T. Galerkin approximations for the two-point boundary value problem using continuous, piecewise polynomial spaces. *Numer. Math.* 1974; 22:99–109.

Douglas Jr J, Dupont T and Wheeler MF. An L^∞ estimate and a superconvergence result for a Galerkin method for elliptic equations based on tensor products of piecewise polynomials. *RAIRO Modél. Math. Anal. Numér.* 1974; 8:61–66.

Douglas Jr J, Dupont T, Percell P and Scott R. A family of C^1 finite elements with optimal approximation properties for various Galerkin methods for 2nd and 4th order problems. *RAIRO Modél. Math. Anal. Numér.* 1979; 13:227–255.

Dupont T and Scott R. Polynomial approximation of functions in Sobolev spaces. *Math. Comput.* 1980; 34:441–463.

Duvaut G and Lions JL. *Inequalities in Mechanics and Physics*. Springer-Verlag, Berlin, 1976.

Eriksson K and Johnson C. Adaptive finite element methods for parabolic problems. I. A linear model problem. *SIAM J. Numer. Anal.* 1991; 28:43–77.

Eriksson K, Estep D, Hansbo P and Johnson C. Introduction to adaptive methods for differential equations. *Acta Numer.* 1995; 4:105–158.

Friedrichs KO. On the boundary value problems of the theory of elasticity and Korn's inequality. *Ann. Math.* 1947; 48:441–471.

Gastaldi L and Nochetto RH. Optimal L^∞ -error estimates for nonconforming and mixed finite element methods of lowest order. *Numer. Math.* 1987; 50:587–611.

Gilbarg D and Trudinger NS. *Elliptic Partial Differential Equations of Second Order* (2nd edn). Springer-Verlag, Berlin, 1983.

Girault V and Scott LR. Hermite interpolation of nonsmooth functions preserving boundary conditions. *Math. Comput.* 2002; 71:1043–1074.

Grisvard P. *Elliptic Problems in Nonsmooth Domains*. Pitman, Boston, 1985.

Hughes TJR. *The Finite Element Method. Linear Static and Dynamic Finite Element Analysis*. Prentice Hall, Englewood Cliffs, 1987 (Reprinted by Dover Publications, New York, 2000).

Jamet P. Estimations d'erreur pour des éléments finis droits presque dégénérés. *RAIRO Anal. Numér.* 1976; 10:43–61.

Kozlov VA, Maz'ya VG and Rossmann J. *Elliptic Boundary Value Problems in Domains with Point Singularities*. American Mathematical Society, 1997.

Kozlov VA, Maz'ya VG and Rossmann J. *Spectral Problems Associated with Corner Singularities of Solutions to Elliptic Problems*. American Mathematical Society, 2001.

Křížek M and Neittaanmäki P. On superconvergence techniques. *Acta Appl. Math.* 1987; 9:175–198.

Křížek M, Neittaanmäki P and Stenberg R (eds). *Finite Element Methods: Superconvergence, Post-processing and A Posteriori*

Estimates, Proc. Conf. Univ. of Jyväskylä, 1996, Lecture Notes in Pure and Applied Mathematics 196. Marcel Dekker, New York, 1998.

Ladeveze P and Leguillon D. Error estimate procedure in the finite element method and applications. *SIAM J. Numer. Anal.* 1983; 20:485–509.

Lenoir M. Optimal isoparametric finite elements and error estimates for domains involving curved boundaries. *SIAM J. Numer. Anal.* 1986; 23:562–580.

Lin Q and Yan NN. *Construction and Analysis of Efficient Finite Element Methods*. Hebei University Press, Baoding, 1996 (in Chinese).

Mansfield L. Approximation of the boundary in the finite element solution of fourth order problems. *SIAM J. Numer. Anal.* 1978; 15:568–579.

Malkus DS and Hughes TJR. Mixed finite element methods-reduced and selective integration techniques: A unification of concepts. *Comput. Methods Appl. Mech. Eng.* 1978; 15:63–81.

Morin P, Nochetto RH and Siebert KG. Local problems on stars: a posteriori error estimation, convergence, and performance. *Math. Comput.* 2003a; 72:1067–1097.

Morin P, Nochetto RH and Siebert KG. Convergence of adaptive finite element methods. *SIAM Rev.* 2003b; 44:631–658.

Natterer F. Über die punktweise Konvergenz finiter Elemente. *Numer. Math.* 1975; 25:67–77.

Nazarov SA and Plamenevsky BA. *Elliptic Problems in Domains with Piecewise Smooth Boundaries*. Walter De Gruyter, Berlin, 1994.

Nečas J. *Les Méthodes Directes en Théorie des Équations Elliptiques*. Masson, Paris, 1967.

Nicaise S. *Polygonal Interface Problems*. Peter D Lang Verlag, Frankfurt am Main, 1993.

Nitsche JA. L^∞ -convergence of finite element approximations. In *Mathematical Aspects of Finite Element Methods*, Lecture Notes in Mathematics 606. Springer-Verlag: New York, 1977; 261–274.

Nitsche JA. On Korn's second inequality. *RAIRO Anal. Numér.* 1981; 15:237–248.

Nitsche JA and Schatz AH. Interior estimates for Ritz-Galerkin methods. *Math. Comput.* 1974; 28:937–958.

Nochetto RH. Removing the saturation assumption in a posteriori error analysis. *Istit. Lombardo Accad. Sci. Lett. Rend. A* 1993; 127:67–82.

Nochetto RH and Wahlbin LB. Positivity preserving finite element approximation. *Math. Comput.* 2002; 71:1405–1419.

Oden JT and Demkowicz LF. *Applied Functional Analysis*. CRC Press, Boca Raton, 1996.

Oden JT and Reddy JN. *An Introduction to the Mathematical Theory of Finite Elements*. Wiley, New York, 1976.

Rannacher R and Scott R. Some optimal error estimates for piecewise linear finite element approximations. *Math. Comput.* 1982; 38:437–445.

Reddy JN. *Applied Functional Analysis and Variational Methods in Engineering*. McGraw-Hill, New York, 1986.

Rivara MC. Algorithms for refining triangular grids suitable for adaptive and multigrid techniques. *Int. J. Numer. Methods Eng.* 1984; 20:745–756.

Rodriguez R. Some remarks on Zienkiewicz-Zhu estimator. *Numer. Methods Partial Diff. Equations* 1994a; 10:625–635.

Rodriguez R. A posteriori error analysis in the finite element method. In *Lecture Notes in Pure and Applied Mathematics*, vol. 164. Marcel Dekker: New York, 1994b; 389–397.

Schatz AH. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comput.* 1974; 28:959–962.

Schatz AH. Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part I. Global estimates. *Math. Comput.* 1998; 67:877–899.

Schatz AH. Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: Part II. Interior estimates. *SIAM J. Numer. Anal.* 2000; 38:1269–1293.

Schatz AH and Wahlbin LB. Maximum norm estimates in the finite element method on plane polygonal domains. Part 1. *Math. Comput.* 1978; 32:73–109.

Schatz AH and Wahlbin LB. Maximum norm estimates in the finite element method on plane polygonal domains. Part 2. *Math. Comput.* 1979; 33:465–492.

Schatz AH and Wahlbin LB. On the quasi-optimality in L_∞ of the \tilde{H}^1 -projection into finite element spaces. *Math. Comput.* 1982; 38:1–22.

Schatz AH and Wahlbin LB. Interior maximum-norm estimates for finite element methods, Part II. *Math. Comput.* 1995; 64:907–928.

Schatz AH, Sloan IH and Wahlbin LB. Superconvergence in finite element methods and meshes that are locally symmetric with respect to a point. *SIAM J. Numer. Anal.* 1996; 33:505–521.

Schatz AH, Thomée V and Wendland WL. *Mathematical Theory of Finite and Boundary Element Methods*. Birkhäuser Verlag, Basel, 1990.

Scott R. *Finite Element Techniques for Curved Boundaries*. Doctoral thesis, Massachusetts Institute of Technology, 1973.

Scott R. Interpolated boundary conditions in the finite element method. *SIAM J. Numer. Anal.* 1975; 12:404–427.

Scott R. Optimal L^∞ estimates for the finite element method on irregular meshes. *Math. Comput.* 1976; 30:681–697.

Scott LR and Zhang S. Finite element interpolation of non-smooth functions satisfying boundary conditions. *Math. Comput.* 1990; 54:485–493.

Strang G and Berger A. The change in solution due to change in domain. *Proceedings AMS Symposium on Partial Differential Equations*. American Mathematical Society: Providence, 1971; 199–205.

Strang G and Fix GJ. *An Analysis of the Finite Element Method*. Prentice Hall, Englewood Cliffs, 1973 (Reprinted by Wellesley-Cambridge Press, Wellesley, 1988).

Szabó BA and Babuška I. *Finite Element Analysis*. John Wiley & Sons, New York, 1991.

Triebel H. *Interpolation Theory, Function Spaces, Differential Operators*. North Holland, Amsterdam, 1978.

Verfürth R. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley-Teubner, New York, 1996.

Verfürth R. A note on polynomial approximation in Sobolev spaces. *Modél. Math. Anal. Numér.* 1999; 33:715–719.

Wahlbin LB. On the sharpness of certain local estimates for \tilde{H}^1 projections into finite element spaces: Influence of a reentrant corner. *Math. Comput.* 1984; 42:1–8.

Wahlbin LB. Local behavior in finite element methods. In *Handbook of Numerical Analysis*, vol. II, Ciarlet FG, Lions JL (eds). North Holland: Amsterdam, 1991; 355–522.

Wahlbin LB. *Superconvergence in Galerkin Finite Element Methods*. Lecture Notes in Mathematics 1605. Springer-Verlag, Berlin, 1995.

Wheeler MF. An optimal L_∞ error estimate for Galerkin approximations to solutions of two-point boundary value problems. *SIAM J. Numer. Anal.* 1973; 1:914–917.

Wloka J. *Partial Differential Equations*. Cambridge University Press, Cambridge, 1987.

Yosida K. *Functional Analysis*, Classics in Mathematics. Springer-Verlag, Berlin, 1995.

Zienkiewicz OC and Taylor RL. *The Finite Element Method* (5th edn). Butterworth-Heinemann, Oxford, 2000.

Zlámal M. Curved elements in the finite element method. I. *SIAM J. Numer. Anal.* 1973; 10:229–240.

Zlámal M. Curved elements in the finite element method. II. *SIAM J. Numer. Anal.* 1974; 11:347–362.

Ženíšek A. Maximum-angle condition and triangular finite element of Hermite type. *Math. Comput.* 1995; 64:929–941.

Chapter 5
The p -version of the Finite Element Method

Barna Szabó¹, Alexander Düster² and Ernst Rank²

¹ Washington University, St. Louis, MO, USA
² Lehrstuhl für Bauinformatik, Technische Universität München, Munich, Germany

| | |
|--------------------------------------|-----|
| 1 Introduction | 119 |
| 2 Implementation | 120 |
| 3 Convergence Characteristics | 126 |
| 4 Performance Characteristics | 131 |
| 5 Applications to Nonlinear Problems | 133 |
| 6 Outlook | 136 |
| Acknowledgments | 137 |
| Notes | 137 |
| References | 137 |
| Further Reading | 139 |

1 INTRODUCTION

The p -version of the finite element method (FEM) is presented as a method for obtaining approximate solutions to generalized formulations of the form

‘Find $u \in X$ such that $B(u, v) = \mathcal{F}(v)$ for all $v \in Y$ ’ (1)

where u and v are scalar or vector functions in one, two, or three dimensions. In the displacement formulation of solid mechanics problems, for example, u is the displacement function, X is the space of admissible displacement functions, v is the virtual displacement function, Y is the space

of admissible virtual displacement functions, $B(u, v)$ is the virtual work of internal stresses, and $\mathcal{F}(v)$ is the virtual work of external forces.

More generally, u (resp. X) is called the *trial function* (resp. *trial space*) and v (resp. Y) is called the *test function* (resp. *test space*), $B(u, v)$ is a bilinear form defined on $X \times Y$ and $\mathcal{F}(v)$ is a linear functional defined on Y . Associated with the spaces X and Y are the norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. The definitive properties of bilinear forms and linear functionals are listed, for example, in Szabó and Babuška (1991), Schwab (1998), and Babuška and Strouboulis (2001).

The definitions for $B(u, v)$, $\mathcal{F}(v)$, X , and Y depend on the choice of the generalized formulation and the boundary conditions. The solution domain will be denoted by Ω and the set of functions u that satisfy the condition $B(u, u) \leq C < \infty$ on Ω will be called the *energy space* and denoted by $E(\Omega)$. The exact solution will be denoted by u_{EX} . The energy norm defined by

$$\|u\|_{E(\Omega)} := \sqrt{\frac{1}{2}B(u, u)} \tag{2}$$

will be associated with the spaces $X \subset E(\Omega)$ and $Y \subset E(\Omega)$. It can be shown that this formulation is equivalent to the minimization of the potential energy functional defined by

$$\Pi(u) := \frac{1}{2}B(u, u) - \mathcal{F}(u) \tag{3}$$

The exact solution u_{EX} of equation (1) is the minimizer of $\Pi(u)$ on the space $X \subset E(\Omega)$.

In the finite element method, finite dimensional subspaces $\tilde{X} \subset X$ and $\tilde{Y} \subset Y$ are constructed. These spaces are characterized by the finite element mesh, the polynomial degrees assigned to the elements, and the mapping

Encyclopedia of Computational Mechanics, Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes. Volume 1: *Fundamentals*. © 2004 John Wiley & Sons, Ltd. ISBN: 0-470-84699-2.

functions. Details are given in Section 2. An approximation to u_{EX} , denoted by u_{FE} , is obtained by solving the finite dimensional problem:

'Find $u_{FE} \in S$ such that $B(u_{FE}, v) = F(v)$ for all $v \in V$ ' (4)

The dimension of V is the number of degrees of freedom, denoted by N .

A key theorem states that the finite element solution u_{FE} minimizes the error in energy norm

$$\|u_{EX} - u_{FE}\|_{E(\Omega)} = \min_{u \in S} \|u_{EX} - u\|_{E(\Omega)} \quad (5)$$

It is seen that the error in energy norm depends on the choice of S . Proper choice of S depends on the regularity of u_{EX} , the objectives of computation, and the desired level of precision.

Another important theorem establishes the following relationship between the error measured in energy norm and the potential energy:

$$\|u_{EX} - u_{FE}\|_{E(\Omega)}^2 = \Pi(u_{FE}) - \Pi(u_{EX}) \quad (6)$$

Proofs are available in Szabó and Babuška (1991). In the p -version, this theorem is used in a posteriori estimation of error in energy norm.

The data of interest are functionals of u_{EX} : $\Psi_1(u_{EX})$, $\Psi_2(u_{EX})$, ... approximated by $\Psi_1(u_{FE})$, $\Psi_2(u_{FE})$, ... An important objective of finite element computations is to establish that the relative errors in the data of interest are small. Therefore, it is necessary to show that

$$|\Psi_i(u_{EX}) - \Psi_i(u_{FE})| \leq \tau_i |\Psi_i(u_{EX})| \quad i = 1, 2, \dots \quad (7)$$

where τ_i are prescribed tolerances. Of course, $\Psi_i(u_{EX})$ is generally unknown; however, $\Psi_i(u_{FE})$ is known to be independent of the choice of the space S . Therefore, if we compute a sequence of finite element solutions corresponding to a hierarchy of spaces $S_1 \subset S_2 \subset S_3 \subset \dots$ then $\Psi_i(u_{FE}) \rightarrow \Psi_i(u_{EX})$. The limiting value of $\Psi_i(u_{FE})$ and hence τ_i can be estimated. The p -version of the finite element method is well suited for the creation of hierarchic finite element spaces and hence the estimation and control of errors in terms of the data of interest.

2 IMPLEMENTATION

From the theoretical point of view, the quality of approximation is completely determined by the finite element space characterized by the finite element mesh Δ , the polynomial

degrees of elements p , and the mapping functions Q (see Section 2.4). Specifically, the finite element space S is a set of functions constructed from polynomials defined on standard elements that are mapped onto the elements of the finite element mesh, subject to the appropriate continuity requirements to ensure that it is a subset of the energy space

$$S := \{u|u \in E(\Omega), u(Q^k) \in S^{p_k}, k = 1, 2, \dots, M(\Delta)\}$$

where Q^k is the mapping function for the k th element, S^{p_k} is the polynomial space of degree p_k associated with the k th element, and $M(\Delta)$ is the number of elements. Different sets of basis functions, called *shape functions*, can be chosen to define the same finite element space; however, there are some important considerations:

1. For a wide range of mapping parameters, the round-off error accumulation with respect to increasing polynomial degree should be as small as possible. (Ideally, the element-level stiffness matrices should be perfectly diagonal, but it is neither necessary nor practical to choose the shape functions in that way in two and three dimensions.)
2. The shape functions should permit computation of the stiffness matrices and load vectors as efficiently as possible.
3. The shape functions should permit efficient enforcement of exact and minimal continuity.
4. The choice of the shape functions affects the performance of iterative solution procedures. For large problems, this can be the dominant consideration.

The first three points suggest that shape functions should be constructed from polynomial functions that have certain orthogonality properties; should be *hierarchic*, that is, the set of shape functions of polynomial degree p should be in the set of shape functions of polynomial degree $p+1$, and the number of shape functions that do not vanish at vertices, edges, and faces should be the smallest possible. Some of the shape functions used in various implementations of the p -version are described in the following.

2.1 Hierarchic shape functions for one-dimensional problems

The classical finite element *nodal basis functions* in one dimension on the standard element $\Omega_e = (-1, 1)$ are illustrated on the left-hand side of Figure 1.

The standard shape functions are defined by the set of Lagrange polynomials

$$N_i^p(\xi) = \prod_{j=1, j \neq i}^{p+1} \frac{\xi - \xi_j}{\xi_i - \xi_j} \quad (8)$$

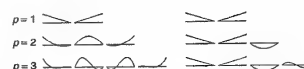


Figure 1. Set of one-dimensional standard and hierarchic shape functions for $p = 1, 2, 3$. (Reproduced by permission of John Wiley & Sons, Ltd from A. Düster, H. Bröker and E. Rank, *Int. J. Numer. Meth. Eng.*, 52, 673–703 (2001).)

The points ξ_j where

$$N_i^p(\xi_j) = \delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \quad (9)$$

are called *nodes*. There are certain advantages in selecting the nodes to be the Gauss–Lobatto points as done in the spectral element method, which is also addressed in this encyclopedia (see Chapter 3, Volume 3). This approach has been modified to suit the p -version of the finite element method in Melenk, Gerdes and Schwab (2001). Note that the sum of all Lagrange polynomials for a given polynomial degree p equals unity:

$$\sum_{i=1}^{p+1} N_i^p(\xi) = 1 \quad (10)$$

Every function that can be represented as a linear combination of this standard basis can be represented also by the set of hierarchic basis functions (see the right-hand side of Figure 1). A principal difference between the two bases is that in the hierarchic case all lower-order shape functions are contained in the higher-order basis. The set of one-dimensional hierarchic shape functions, introduced by Szabó and Babuška (1991), is given by

$$N_1(\xi) = \frac{1}{2}(1 - \xi) \quad (11)$$

$$N_2(\xi) = \frac{1}{2}(1 + \xi) \quad (12)$$

$$N_i(\xi) = \phi_{i-1}(\xi), \quad i = 3, 4, \dots, p+1 \quad (13)$$

with

$$\begin{aligned} \phi_j(\xi) &= \sqrt{\frac{2j-1}{2}} \int_{-1}^{\xi} L_{j-1}(x) dx \\ &= \frac{1}{\sqrt{4j-2}} (L_j(\xi) - L_{j-2}(\xi)), \quad j = 2, 3, \dots \quad (14) \end{aligned}$$

where $L_j(\xi)$ are the Legendre polynomials. The first two shape functions $N_1(\xi), N_2(\xi)$ are called *nodal shape functions* or *nodal modes*. Because

$$N_i(-1) = N_i(1) = 0, \quad i \geq 3 \quad (15)$$

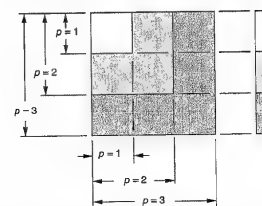


Figure 2. Hierarchic structure of stiffness matrix and load vector with $p = 3$. (Reproduced by permission of John Wiley & Sons, Ltd from E. Stein (Editor), *Error-controlled Adaptive Finite Elements in Solid Mechanics*, 263–307 (2002).)

the functions $N_i(\xi)$, $i = 3, 4, \dots$ are called *internal shape functions*, *internal modes*, or *bubble modes*. The orthogonality property of Legendre polynomials implies

$$\int_{-1}^1 \frac{dN_i}{d\xi} \frac{dN_j}{d\xi} d\xi = \delta_{ij}, \quad i \geq 3 \quad \text{and} \quad j \geq 1 \quad (16)$$

If equations are ordered in such a way that all linear modes are numbered from 1 to n_1 , all quadratic modes are numbered from $n_1 + 1$ to n_2 and so on, stiffness matrices corresponding to polynomial order 1 to $p-1$ are submatrices of the stiffness matrix corresponding to polynomial order p . Figure 2 depicts the structure of a stiffness matrix and a load vector corresponding to polynomial degree of $p = 3$ schematically.

2.2 Hierarchic shape functions for quadrilaterals

The standard quadrilateral finite element is shown in Figure 4. Two types of standard polynomial spaces, the *trunk space* $S_0^{p,p}(\Omega_e)$ and the *tensor product space* $S_0^{p,p}(\Omega_e)$, are discussed in the following. The *tensor product space* $S_0^{p,p}(\Omega_e)$ consists of all polynomials on $\Omega_e = [(-1, 1) \times (-1, 1)]$ spanned by the set of monomials $\xi^i \eta^j$ where $i = 0, 1, \dots, p_\xi$, $j = 0, 1, \dots, p_\eta$, whereas the *trunk space* $S_0^{p,p}(\Omega_e)$ on $\Omega_e = [(-1, 1) \times (-1, 1)]$ is spanned by the set of all monomials

- $\xi^i \eta^j$ with $i = 0, \dots, p_\xi$, $j = 0, \dots, p_\eta$, $i + j = 0, \dots, \max(p_\xi, p_\eta)$
- $\xi \eta$ for $p_\xi = p_\eta = 1$
- $\xi^2 \eta$ for $p_\xi \geq 2$
- $\xi \eta^2$ for $p_\eta \geq 2$

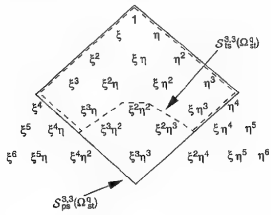


Figure 3. The trunk space $S_0^{3,3}(\Omega_4^2)$ and the tensor product space $S_{pq}^{3,3}(\Omega_4^2)$. (Reproduced by permission of John Wiley & Sons, Ltd from E. Stein (Editor), *Error-controlled Adaptive Finite Elements in Solid Mechanics*, 263–307 (2002).)

The difference between the two standard polynomial spaces can be readily visualized when considering the spanning sets in Pascal's triangle. The set of monomials for $p_k = p_\eta = 3$ for both the trunk and the tensor product space is shown in Figure 3. All monomials inside the dashed line span the trunk space $S_0^{3,3}(\Omega_4^2)$, whereas the monomials bordered by the solid line are essential for the tensor product space $S_{pq}^{3,3}(\Omega_4^2)$.

Two-dimensional shape functions can be classified into three groups: nodal, edge, and internal shape functions. Using the numbering convention shown in Figure 4, these shape functions are described in the following.

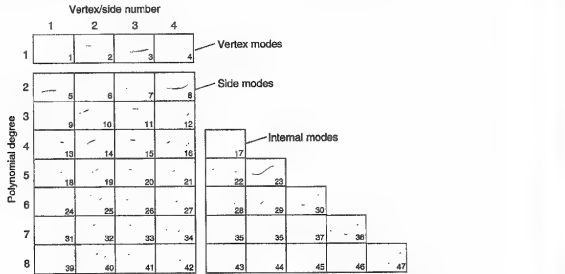


Figure 5. Hierarchic shape functions for quadrilateral elements. Trunk space, $p = 1$ to $p = 8$. (From *Finite Element Analysis*, B. Szabó and I. Babuška; Copyright (1991) John Wiley & Sons, Inc. This material is used by permission of John Wiley & Sons, Inc.)

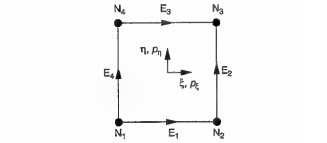


Figure 4. Standard quadrilateral element Ω_4^2 : definition of nodes, edges, and polynomial degree.

1. **Nodal or vertex modes:** The nodal modes

$$N_{i,1}^{N_i}(\xi, \eta) = \frac{1}{4}(1 + \xi_i \xi)(1 + \eta_i \eta), \quad i = 1, \dots, 4 \quad (17)$$

are the standard bilinear shape functions, well known from the isoparametric four-noded quadrilateral element. (ξ_i, η_i) denote the local coordinates of the i th node.

2. **Edge or side modes:** These modes are defined separately for each individual edge, they vanish at all other edges. The corresponding modes for edge E_1 read:

$$N_{i,1}^{E_1}(\xi, \eta) = \frac{1}{2}(1 - \eta)\phi_i(\xi), \quad i \geq 2 \quad (18)$$

3. **Internal modes:** The internal modes

$$N_{i,j}^{int}(\xi, \eta) = \phi_i(\xi)\phi_j(\eta), \quad i, j \geq 2 \quad (19)$$

are purely local and vanish at the edges of the quadrilateral element.

As already indicated, the indices i, j of the shape functions denote the polynomial degrees in the local directions ξ, η . In Figure 5, all hierarchic shape functions that span the trunk space are plotted up to order $p = 8$.

2.3 Hierarchic shape functions for hexahedra

The implementation of high-order finite elements in three dimensions can be based on a hexahedral element formulation (see Figure 6), again using the hierarchic shape functions introduced by Szabó and Babuška (1991). High-order hexahedral elements are suited for solid, 'thick' structures and for thin-walled structures alike. In the case of plate- or shell-like structures, one local variable can be identified to correspond with the thickness direction and it is possible to choose the polynomial degree in the thickness direction differently from those in the in-plane direction; see Düster, Bröker and Rank (2001). Generalizing the two-dimensional concept, three-dimensional shape functions can be classified into four groups:

1. **Nodal or vertex modes:** The nodal modes

$$N_{i,j,k}^{N_i}(\xi, \eta, \zeta) = \frac{1}{8}(1 + \xi_i \xi)(1 + \eta_j \eta)(1 + \zeta_k \zeta), \quad i = 1, \dots, 8 \quad (20)$$

are the standard trilinear shape functions, well known from the isoparametric eight-noded brick element. (ξ_i, η_j, ζ_k) are the local coordinates of the i th node.

2. **Edge modes:** These modes are defined separately for each edge. If we consider, for example, edge E_1 (see Figure 6), the corresponding edge modes read:

$$N_{i,j,1}^{E_1}(\xi, \eta, \zeta) = \frac{1}{4}(1 - \eta)(1 - \zeta)\phi_i(\xi), \quad i \geq 2 \quad (21)$$

3. **Face modes:** These modes are defined separately for each individual face. If we consider, for example, face F_1 , the corresponding face modes read:

$$N_{i,j,1}^{F_1}(\xi, \eta, \zeta) = \frac{1}{2}(1 - \zeta)\phi_i(\xi)\phi_j(\eta), \quad i, j \geq 2 \quad (22)$$

4. **Internal modes:** The internal modes

$$N_{i,j,k}^{int}(\xi, \eta, \zeta) = \phi_i(\xi)\phi_j(\eta)\phi_k(\zeta), \quad i, j, k \geq 2 \quad (23)$$

are purely local and vanish at the faces of the hexahedral element.

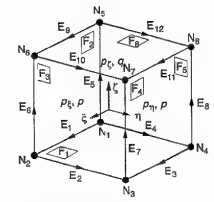


Figure 6. Standard hexahedral element Ω_8^3 : definition of nodes, edges, faces, and polynomial degree.

The indices i, j, k of the shape functions denote the polynomial degrees in the local directions ξ, η, ζ .

Three different types of trial spaces can be readily defined: the *trunk space* $S_0^{p_\xi, p_\eta, p_\zeta}(\Omega_8^3)$, the *tensor product space* $S_{pq}^{p_\xi, p_\eta, p_\zeta}(\Omega_8^3)$, and an *anisotropic tensor product space* $S^{p_\xi, p_\eta, p_\zeta}(\Omega_8^3)$. A detailed description of these trial spaces can be found in Szabó and Babuška (1991). The polynomial degree for the trial spaces $S_0^{p_\xi, p_\eta, p_\zeta}(\Omega_8^3)$ and $S_{pq}^{p_\xi, p_\eta, p_\zeta}(\Omega_8^3)$ can be varied separately in each local direction (see Figure 6). Differences between the trunk and product spaces occur in the face modes and internal modes only. For explanation, we first consider the face modes, for example, the modes for face 1. Indices i, j denote the polynomial degrees of the face modes in ξ and η direction, respectively.

Face modes (face F_1): $N_{i,j,1}^{F_1}(\xi, \eta, \zeta) = 1/2(1 - \zeta)\phi_i(\xi)\phi_j(\eta)$

| trunk space | tensor product space |
|---|------------------------|
| $i = 2, \dots, p_\xi - 2$ | $i = 2, \dots, p_\xi$ |
| $j = 2, \dots, p_\eta - 2$ | $j = 2, \dots, p_\eta$ |
| $i + j = 4, \dots, \max\{p_\xi, p_\eta\}$ | |

The definition of the set of internal modes is very similar. Indices i, j, k now denote the polynomial degrees in the three local directions ξ, η , and ζ .

Internal modes: $N_{i,j,k}^{int}(\xi, \eta, \zeta) = \phi_i(\xi)\phi_j(\eta)\phi_k(\zeta)$

| trunk space | tensor product space |
|--|-------------------------|
| $i = 2, \dots, p_\xi - 4$ | $i = 2, \dots, p_\xi$ |
| $j = 2, \dots, p_\eta - 4$ | $j = 2, \dots, p_\eta$ |
| $k = 2, \dots, p_\zeta - 4$ | $k = 2, \dots, p_\zeta$ |
| $i + j + k = 6, \dots, \max\{p_\xi, p_\eta, p_\zeta\}$ | |

The space $S^{p,p,q}(\Omega_h^b)$ defines an anisotropic set of shape functions determined by two polynomial degrees p and q (see Figure 6). All shape functions of higher order in ξ and η direction are associated with the polynomial degree p . These shape functions correspond to the edges 1, 2, 3, 4, 9, 10, 11, 12, to the faces 1 and 6 and to all internal modes. Shape functions for faces 1 and 6 are equal to those of the trunk space $S_{tr}^{p_1,p_1,p_1}(\Omega_h^b)$ with $p_1 = p_1 = p$. q defines the degree of all shape functions of higher order in ζ -direction that are associated with the edges 5, 6, 7, 8, with the faces 2, 3, 4, 5, and with all internal modes. The modes corresponding to the faces 2, 3, 4, 5, are equal to those of the tensor product space $S_{tp}^{p_1,p_1,p_1}(\Omega_h^b)$ with $p = p_1 = p_1$ and $q = p_1$. Considering a polynomial degree $p = q = p_1 = p_1 = p_1$, one observes that the number of internal modes of $S^{p,p,q}(\Omega_h^b)$ is larger than that of the trunk space $S_{tr}^{p_1,p_1,p_1}(\Omega_h^b)$ but smaller than that of the tensor product space $S_{tp}^{p_1,p_1,p_1}(\Omega_h^b)$.

Owing to the built-in anisotropic behavior of the trial space $S^{p,p,q}(\Omega_h^b)$, it is important to consider the orientation of the local coordinates of a hexahedral element. Figure 7 shows how hexahedral elements should be oriented when three-dimensional, thin-walled structures are discretized. The local coordinate ζ of the hexahedral element corresponds to the thickness direction. If the orientation of all elements is the same then it is possible to construct discretizations where the polynomial degree for the in-plane and thickness directions of thin-walled structures can be treated differently.

2.4 Mapping

In low-order finite element analysis (FEA), the most frequently used mapping technique for the geometric description of the domain of computation is the application of isoparametric elements where the standard shape functions are used for the geometric description of elements. The same shape functions are used for the approximation of the unknown solution and for the shape of the elements.

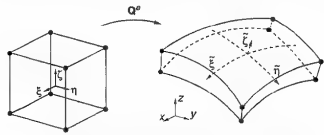


Figure 7. Modelling thin-walled structures with hexahedral elements.

Using elements of order $p = 1$ or $p = 2$, the boundary of the domain is approximated by a polygonal or by a piecewise parabolic curve, respectively. As the mesh is refined, the boundary of the domain is approximated more and more accurately. When using the p -version, on the other hand, the mesh remains fixed. It is therefore important to model the geometry of the structure accurately with the fixed number of elements. This calls for a method that is able to describe complex geometries using only a few elements. Gordon and Hall (1973a,b) proposed the *blending function method* that is usually applied when describing curved boundaries of p -version finite elements; see, for example, Szabó and Babuška (1991) and Düster, Bröcker and Rank (2001). After introducing blending function mapping, an example will compare polynomial interpolation versus exact blending mapping and demonstrate the necessity of a precise description of geometry.

2.4.1 The blending function method

Consider a quadrilateral element as shown in Figure 8 where edge E_2 is assumed to be part of a curved boundary. The shape of edge E_2 is assumed to be defined by a parametric function $E_2 = [E_{2x}(\eta), E_{2y}(\eta)]^T$, where η is the local coordinate of the element. The transformation of the local coordinates $\xi = [\xi, \eta]^T$ into the global coordinates $\mathbf{x} = [x, y]^T = \mathbf{Q}^p = [Q_x^p(\xi, \eta), Q_y^p(\xi, \eta)]^T$ can be formulated by the two functions

$$\begin{aligned} x &= Q_x^p(\xi, \eta) = \sum_{i=1}^4 N_{i,1}^p(\xi, \eta) X_i \\ &\quad + \left(E_{2x}(\eta) - \left(\frac{1-\eta}{2} X_2 + \frac{1+\eta}{2} X_3 \right) \right) \frac{1+\xi}{2} \\ y &= Q_y^p(\xi, \eta) = \sum_{i=1}^4 N_{i,1}^p(\xi, \eta) Y_i \\ &\quad + \left(E_{2y}(\eta) - \left(\frac{1-\eta}{2} Y_2 + \frac{1+\eta}{2} Y_3 \right) \right) \frac{1+\xi}{2} \end{aligned} \quad (24)$$

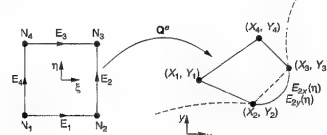


Figure 8. Blending function method for quadrilateral elements.

where the first term corresponds to the standard bilinear mapping that is familiar from the isoparametric concept for quadrilateral elements with $p = 1$. The second term takes the curved edge E_2 into account. Therefore, the bilinear mapping is augmented by the blended difference between the curve $E_2 = [E_{2x}(\eta), E_{2y}(\eta)]^T$ and the straight line connecting the nodes N_2 and N_3 . The blending term $(1+\xi)/2$ ensures that the opposite edge E_4 – where $(1+\xi)/2 = 0$ – is not affected by the curvilinear description of edge E_2 .

If a quadrilateral in which all edges are curved is to be considered, the blending function method can be expanded such that the mapping reads

$$\begin{aligned} x &= Q_x^p(\xi, \eta) = \frac{1}{2}(1-\eta)E_{1x}(\xi) + \frac{1}{2}(1+\xi)E_{2x}(\eta) \\ &\quad + \frac{1}{2}(1+\eta)E_{3x}(\xi) + \frac{1}{2}(1-\xi)E_{4x}(\eta) \\ &\quad - \sum_{i=1}^4 N_{i,1}^p(\xi, \eta) X_i \\ y &= Q_y^p(\xi, \eta) = \frac{1}{2}(1-\eta)E_{1y}(\xi) + \frac{1}{2}(1+\xi)E_{2y}(\eta) \\ &\quad + \frac{1}{2}(1+\eta)E_{3y}(\xi) + \frac{1}{2}(1-\xi)E_{4y}(\eta) \\ &\quad - \sum_{i=1}^4 N_{i,1}^p(\xi, \eta) Y_i \end{aligned} \quad (25)$$

where

$$\begin{aligned} E_{ix}(\xi), E_{iy}(\xi), \quad &\text{for } i = 1, 3 \\ E_{ix}(\eta), E_{iy}(\eta), \quad &\text{for } i = 2, 4 \end{aligned} \quad (26)$$

are parametric functions describing the shape of the edges E_i , $i = 1, 2, 3, 4$. Therefore the blending function method allows arbitrary parametric descriptions of the edges of elements.

2.4.2 Accuracy of mapping versus polynomial interpolation

The following numerical example demonstrates the importance of accurate representation of geometry when a p -extension is to be applied in order to find a finite element approximation. A quarter of a linear elastic plate with a central circular hole and unit thickness (1 mm) is loaded by a traction $T_x = 100$ MPa (see Figure 9). The dimensions are chosen to be $b = h = 100$ mm and $R = 10$ mm. At the lower and right side of the plate, symmetry conditions are imposed. The isotropic linear elastic material behavior is characterized by Young's modulus $E = 206900$ MPa, Poisson's ratio $\nu = 0.29$, and plane

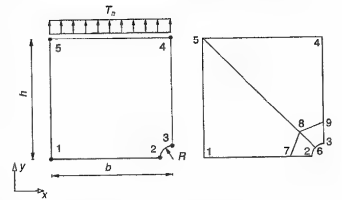


Figure 9. Perforated square plate under uniform tension.

stress assumptions. The strain energy of the plate – obtained by an 'overkill' finite element approximation – amounts to 247.521396 Nmm. The plate is discretized by four quadrilateral elements and the circle with radius $R = 10$ mm is represented by applying

1. **exact blending:** that is, the exact parametric description of a circle is applied;
2. **parabolic description:** two parabolas are used to interpolate the circle with a corresponding relative error $(|R - \tilde{R}|/R)100(\%) < 0.0725(\%)$, where \tilde{R} denotes the radius of the interpolated circle.

A p -extension based on the tensor product space $S_{tp}^{p,p}(\Omega_h^b)$, $p = 1, \dots, 8$ is performed and the relative error in energy norm for both the exact blending and the parabolic boundary interpolation is plotted versus the degrees of freedom on a log-log scale in Figure 10. Owing to the smoothness of the exact solution of the problem, the p -extension in conjunction with the exact blending shows exponential rate of convergence (see equation (29) in Section 3.2). In the case of the parabolic boundary interpolation, the convergence rate of the p -extension deteriorates for $p \geq 3$ and the strain energy finally converges to an incorrect value. Consider the stresses, for instance, stress component σ_{yy} at point 2; we observe that the p -extension with $p = 1, \dots, 20$ and exact blending converges rapidly while the stress obtained with parabolic boundary interpolation diverges (see Figure 11).

Although the relative error of the parabolic geometric interpolation seems to be very small, it has a strong influence on the accuracy of the p -extension. The strain energy of the approximation converges to an incorrect value and the stress component σ_{yy} at point 2 even diverges. The reason for this is that an artificial stress singularity is introduced. Considering the first derivatives at the interelement node 6, and at symmetry nodes 2

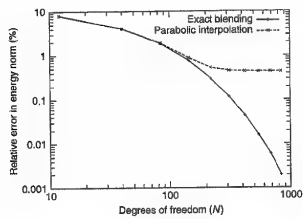


Figure 10. Influence of the blending on the relative error in energy norm.

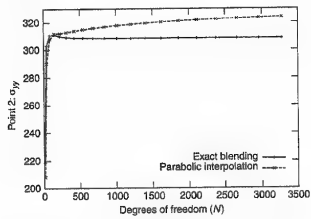


Figure 11. Influence of the blending on the stress component σ_{yy} at point 2.

and 3, discontinuities are observed. They lead to stress singularities similar to stress concentrations at corners. One way of avoiding these stress singularities is to use the exact blending or to apply the so-called quasi-regional mapping described in Királyfalvi and Szabó (1997). The idea of the quasi-regional mapping is to combine the blending function method with a polynomial interpolation of geometry, using optimal collocation points; see Chen and Babuska (1995, 1996). An example of the effectiveness of this kind of mapping is given in connection with a geometrically nonlinear problem in Section 5.2. A detailed comparison of exact and polynomial blending is given by Bröker (2001).

3 CONVERGENCE CHARACTERISTICS

In this section, some key theoretical results that establish relationships between the error in energy norm and the

number of degrees of freedom associated with hierarchic sequences of finite element spaces: $S_1 \subset S_2 \subset \dots$ are presented.

In the early implementations of the finite element method, the polynomial degrees were restricted to $p = 1$ or $p = 2$ only. Finite element spaces were enlarged by mesh refinement, that is, by reducing the diameter of the largest element, denoted by h . Subsequently, this limitation was removed, allowing enlargement of finite element spaces by increasing the polynomial degree of elements, denoted by p , while keeping the mesh fixed. To distinguish between the two approaches, the terms ' h -version' and ' p -version' gained currency. We will consider three strategies for constructing finite element spaces:

- h -Extension:** The polynomial degree of elements is fixed, typically at some low number, such as $p = 1$ or $p = 2$, and the number of elements is increased such that h is progressively reduced.
- p -Extension:** The mesh is fixed and the polynomial degree of elements is increased.
- hp -Extension:** The mesh is refined and the polynomial degrees of elements are concurrently increased.

A fourth strategy, not considered here, introduces basis functions, other than the mapped polynomial basis functions described in Section 2, to represent some local characteristics of the exact solution. This is variously known as the *space enrichment method*, *partition of unity method*, and *meshless method*.

It is of considerable practical interest to know how the first space S_1 should be constructed and when and how h -extension, p -extension, or hp -extension should be used. The underlying principles and practical considerations are summarized in the following.

3.1 Classification

It is useful to establish a simple classification for the exact solution based on a priori information available concerning its regularity. The exact solution, denoted by u_{EX} in the following, may be a scalar function or a vector function.

Category A: u_{EX} is analytic everywhere on the solution domain including the boundaries. By definition, a function is analytic in a point if it can be expanded into a Taylor series about that point. The solution is in category A also when analytical continuation is applicable.

Category B: u_{EX} is analytic everywhere on the solution domain including the boundaries, with the exception of a finite number of points (or in 3D, a finite number of points and edges). The locations where the

exact solution is not analytic are called *singular points* or *singular edges*. The great majority of practical problems in solid mechanics belong in this category. Problems in category B are characterized by *piecewise analytic data*, that is, the domain is bounded by piecewise analytic functions and/or the boundary conditions are piecewise analytic.

Category C: u_{EX} is neither in category A nor in category B.

At corner singularities and at intersections of material interfaces in two-dimensional problems, the exact solution typically can be written in the form

$$u_{EX} = \sum_{i=1}^{\infty} A_i r^{\lambda_i} F_i(\theta), \quad r < \rho, \quad \lambda_{\min} > 0 \quad (27)$$

where r, θ are polar coordinates centered on the singular point, A_i, λ_i are real numbers, F_i is an analytic (or piecewise analytic) vector function, and ρ is the radius of convergence. Additional details can be found in Grisvard (1985). This is known as an asymptotic expansion of the solution in the neighborhood of a singular point. Analogous expressions can be written for one and three dimensions with $\lambda_{\min} > 1 - d/2$ where d is the number of spatial dimensions. The minimum value of λ_i corresponding to a nonzero coefficient A_i characterizes the regularity (also called 'smoothness') of the exact solution. In the following section, the key theorems concerning the asymptotic rates of convergence of the various extension processes are summarized.

3.2 A priori estimates

A priori estimates of the rates of convergence are available for solutions in categories A, B, and C. Convergence is either algebraic or exponential. The algebraic estimate is of the form

$$\|u_{EX} - u_{FE}\|_{E(\Omega)} \leq \frac{k}{N^\beta} \quad (28)$$

and the exponential estimate is of the form

$$\|u_{EX} - u_{FE}\|_{E(\Omega)} \leq \frac{k}{\exp(\gamma N^\delta)} \quad (29)$$

These estimates should be understood to mean that there exists some positive constant k , and a positive constant β (resp. γ and δ) that depend on u_{EX} , such that the error will be bounded by the algebraic (resp. exponential) estimate as the number of degrees of freedom N is increased. These estimates are sharp for sufficiently large N .

The asymptotic rates of convergence for two-dimensional problems are summarized in Table 1 and for three-dimensional problems in Table 2. In these tables, p (resp. λ) represents the minimum polynomial degree assigned to the elements of a finite element mesh (resp. λ_{\min} in equation (27)) (see Chapter 4, this Volume).

3.3 The choice of finite element spaces

The theoretical results described in Section 3.2 provide an important conceptual framework for the construction of finite element spaces (see Chapter 3, this Volume).

3.3.1 Problems in category A

Referring to Tables 1 and 2, it is seen that for problems in category A, exponential rates of convergence are possible through p - and hp -extensions. These convergence rates can be realized provided that all singular points lie on element vertices and edges. For both the p - and hp -extensions, the optimal mesh consists of the smallest number of elements required to partition the solution domain into triangular

Table 1. Asymptotic rates of convergence in two dimensions.

| Category | Type of extension | | |
|----------|--|----------------------------------|----------------------------------|
| | h | p | hp |
| A | Algebraic $\beta = p/2$ | Exponential $\theta \geq 1/2$ | Exponential $\theta \geq 1/2$ |
| B | Algebraic Note 1 $\beta = (1/2) \min(p, \lambda)$ | Algebraic $\beta = \lambda$ | Exponential $\theta \geq 1/3$ |
| C | Algebraic $\beta > 0$ | Algebraic $\beta > 0$ | Note 2 |

Note 1: Uniform or quasi-uniform mesh refinement is assumed. In the case of optimal mesh refinement, $\beta_{\max} = p/2$.

Note 2: When u_{EX} has a recognizable structure, then it is possible to achieve faster than algebraic rates of convergence with hp -adaptive methods.

Table 2. Asymptotic rates of convergence in three dimensions.

| Category | Type of extension | | |
|----------|----------------------------|----------------------------------|----------------------------------|
| | h | p | hp |
| A | Algebraic $\beta = p/3$ | Exponential $\theta \geq 1/3$ | Exponential $\theta \geq 1/3$ |
| B | Note 3 | | |
| C | Algebraic $\beta > 0$ | Algebraic $\beta > 0$ | Note 2 |

Note 3: In three dimensions, u_{EX} cannot be characterized by a single parameter. Nevertheless, the rate of p -convergence is at least twice the rate of h -convergence.

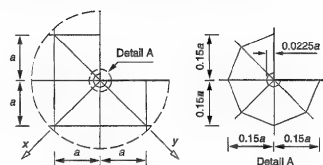


Figure 12. Example of a geometric mesh (detail).

and quadrilateral elements in two dimensions; tetrahedral, pentahedral, and hexahedral elements in three dimensions.

When h -extensions are used, the optimal rate of convergence in 2D is algebraic with $\beta = p/2$. The optimal mesh grading depends on both p and the exact solution.

3.3.2 Problems in category B

When the exact solution can be written in the form of equation (27), there is an optimal design of the discretization in the neighborhood of the singular point. The finite elements should be laid out so that the sizes of elements decrease in geometric progression toward the singular point (located at $x = 0$) and the polynomial degrees of elements increase away from the singular point. The optimal grading is $q = (\sqrt{2} - 1)^2 \approx 0.17$ that is independent of λ_{\min} . In practice, $q = 0.15$ is used. These are called *geometric meshes*. An example of a geometric mesh in two dimensions is given in Figure 12.

The ideal distribution of polynomial degrees is that the lowest polynomial degree is associated with the smallest element and the polynomial degrees increase linearly away from the singular points. This is because the errors in the vicinity of singular points depend primarily on the size of elements, whereas errors associated with elements farther from singular points, where the solution is smooth, depend mainly on the polynomial degree of elements. In practice, uniform p -distribution is used, which yields very nearly optimal results in the sense that convergence is exponential, and the work penalty associated with using uniform polynomial degree distribution is not substantial.

3.4 A simple 1D model problem

In this section, we will consider an axially loaded linear elastic bar as depicted in Figure 13.

Although the solution of the underlying simple model problem (30)–(32) can be stated in a closed form, it

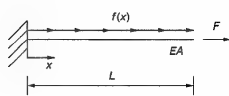


Figure 13. Linear elastic bar.

is worth studying because it implies many of the features that also appear in more complex models. Furthermore, the general concept of the p -version can be readily represented when considering the simple one-dimensional model problem. A tutorial program for a one-dimensional h -, p -, and hp -version of the finite element method, where the following problems can be investigated in detail, has been implemented in Maple [1]. The solution $u(x)$ (length) of the ordinary differential equation (30) describes the displacement of the bar in x -direction, being loaded by a traction $f(x)$ (force/length) and a load F (force). E (force/length²) denotes Young's modulus, A (length²) the cross-sectional area, and L (length) the length of the bar.

$$-(EAu')' = f(x) \quad \text{on } \Omega = [x|0 \leq x \leq L] \quad (30)$$

$$u = 0 \quad \text{at } x = 0 \quad (31)$$

$$EAu' = F \quad \text{at } x = L \quad (32)$$

For the sake of simplicity, it is assumed that the displacement $u(x)$ and strain $\epsilon = du/dx$ are small and that the bar exhibits a linear elastic stress-strain relationship, that is, $\sigma = E\epsilon$ with σ being uniformly distributed over the cross-sectional area A . Equation (31) defines a Dirichlet boundary condition at $x = 0$ and equation (32), a Neumann boundary condition at $x = L$. For a detailed study of this model problem, see Szabó and Babuška (1991). The variational or weak formulation of the model problem (30)–(32), which is the basis for a finite element approximation can be stated as follows:

Find $u \in X$ satisfying (homogeneous) Dirichlet boundary conditions, such that

$$B(u, v) = \mathcal{F}(v) \quad \text{for all } v \in Y \quad (33)$$

where

$$B(u, v) = \int_0^L EAu'v' dx$$

$$\text{and } \mathcal{F}(v) = \int_0^L f v dx + Fv(L) \quad (34)$$

3.4.1 A numerical example with a smooth solution

Figure 13 shows an elastic bar where it is assumed that $EA = L = 1$, $f(x) = -\sin(8x)$ and $F = 0$. The p -version discretizations consist of one element with $p = 1, 2, 3, \dots, 8$, whereas the h -version is based on a uniformly refined mesh with up to eight linear ($p = 1$) elements.

First, we will consider the p -version discretization. The exact solution $u_{EX}(x) = -(1/64)\sin(8x) + (1/8)\cos(8x)$ of the problem (33)–(34) is approximated by a polynomial expression on the basis of the hierarchic shape functions (11)–(13)

$$u_{FE}(\xi) = N_1(\xi)U_1 + N_2(\xi)U_2 + \sum_{p=2}^{p_{\max}} N_{p+1}(\xi)a_{p+1} \quad (35)$$

where $p_{\max} = 8$. U_1 and U_2 denote the nodal displacements, whereas a_3, \dots, a_9 are coefficients determining the higher-order terms of the approximation $u_{FE}(\xi)$. Owing to the orthonormality property (16) of the higher-order shape functions, the element stiffness matrix, $K_{ij}^e = (2/L) \int_{-1}^1 EA(dN_i/d\xi)(dN_j/d\xi) d\xi$, $i, j = 1, 2, 3, \dots, 9$, is almost perfectly diagonal:

$$K^e = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 2 & 0 & \dots & 0 \\ 0 & 0 & 0 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 2 \end{bmatrix} \quad (36)$$

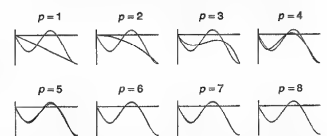
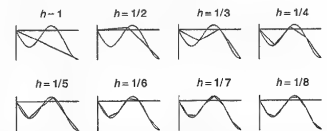
Computing the element load vector, $F_i^e = (L/2) \int_{-1}^1 N_i(\xi) f(x(\xi)) d\xi$, $i = 1, 2, 3, \dots, 9$, one finds

$$F^e = [-0.1095, -0.0336, -0.0269, -0.0714, 0.0811, 0.0433, -0.0230, -0.0073, 0.0026]^T \quad (37)$$

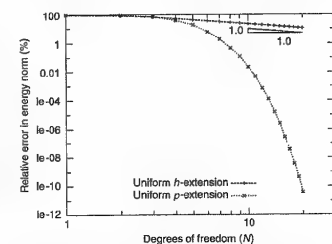
Because of the homogeneous Dirichlet boundary condition ($u(0) = u_{FE}(0) = 0 \rightarrow U_1 = 0$), the solution of the resulting diagonal equation system is trivial in this case. In Figure 14, the p -version approximation $u_{FE}(x)$ for $p = 1, 2, 3, \dots, 8$ is plotted together with the exact solution of the problem. For a first comparison of the accuracy, the same problem is solved by applying the h -version with $p = 1$ based on a uniformly refined mesh with decreasing element size $h_i = 1/i$, $i = 1, \dots, 8$. Again, the approximation and the exact solution is drawn (see Figure 15).

In Figure 16, the relative error in energy norm

$$(\epsilon_r)_{E(\Omega)} = \frac{\|u_{EX} - u_{FE}\|_{E(\Omega)}}{\|u_{EX}\|_{E(\Omega)}} \quad (38)$$

Figure 14. p -version solution $u_{FE}(x)$ based on one element with $p = 1, 2, 3, \dots, 8$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>Figure 15. h -version solution $u_{FE}(x)$ based on a uniform refined mesh with $p = 1$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

is plotted versus the number of degrees of freedom in a double logarithmic style. By the classification given in Section 3.2, this problem is in category A, where the p -version exhibits exponential convergence (29), whereas the asymptotic rate of convergence of the h -extension is algebraic (28). For category A problems in one dimension, the parameter β in equation (28) is $\beta = p$. Since in this case $p = 1$, the asymptotic rate of convergence is 1, as shown in Figure 16.

Figure 16. Comparison of the h - and p -version: relative error in energy norm.

3.4.2 A numerical example with a nonsmooth solution

In the following example, we will again consider the weak formulation (33)–(34) of the model problem (30)–(32) where $f(x)$ is now chosen such that the exact solution is nonsmooth. We define $f(x) = \lambda(\lambda - 1)x^{\lambda-2}$, $F = 0$ and $EA = L = 1$, resulting in an exact solution $u_{EX} = -x^\lambda + \lambda x$, where λ is the parameter controlling the smoothness of the solution. If $\lambda < 1.0$, then the first derivative of the exact solution will exhibit a singularity at $x = 0$ and the given problem will be in category B. Note that $\lambda > 1/2$ is a necessary condition for obtaining a finite strain energy of the exact solution. For the following numerical example, λ is chosen to be 0.65.

In Figure 17, the relative error in energy norm (38) is plotted versus the number of degrees of freedom on a log-log scale. p -Extension was performed on one element with $p = 1, \dots, 50$, whereas the h -extension was performed on meshes with equal sized elements $h = 1, \dots, 1/50$ with $p = 1$. Since the given problem is in category B, both extensions show algebraic convergence of type (28). The asymptotic rate of convergence of the h -extension is given by

$$\beta = \min(p, \lambda - \frac{1}{2}) = 0.15 \quad (39)$$

and can be clearly observed in Figure 17. The rate of convergence of the uniform p -extension is twice the rate of the uniform h -extension. This is due to the fact that the point where the exact solution exhibits singular behavior coincides with a node.

When combining mesh refinement with an increase in polynomial degree, exponential convergence in energy norm (29) can be achieved with an hp -extension, even when

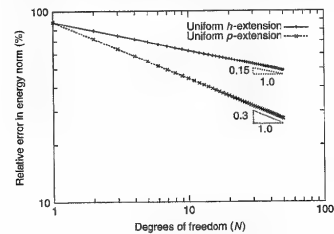


Figure 17. Comparison of the h - and p -version: relative error in energy norm.

the exact solution u_{EX} has singularities. The mesh is refined towards the singular points by geometric progression using the common factor $q = 0.15$. The location of the nodal points X_i is given by

$$X_i = \begin{cases} 0 & \text{for } i = 0 \\ Lq^{n_d-i} & \text{for } i = 1, 2, \dots, n_d \end{cases} \quad (40)$$

A polynomial degree $p_{\max} = 1$ is assigned to the element at the singularity, and increases linearly away from the singular point to the maximum degree

$$p_{\max} = (2\lambda - 1)(n_d - 1) \quad (41)$$

where λ denotes the smoothness of the solution and n_d the total number of elements of the corresponding mesh. With this hp -extension, one obtains an exponential convergence in energy norm as shown in Figure 18 (hp -version, $q = 0.15$, $\lambda = 0.65$). Using about 100 degrees of freedom, the error is by several orders of magnitude smaller than that of a uniform p -version with one element or of a uniform h -version with $p = 1$.

Figure 18 also shows the results of uniform p -extensions obtained on geometrically refined meshes with $q = 0.15$. These extensions are for meshes with $n_d = 4, 8, 12, 16, 20, 24$ elements with p being uniformly increased from 1 to 8. In the preasymptotic range, the p -extension on fixed, geometrically graded meshes shows an exponential convergence rate. In the asymptotic range, the exponential convergence decreases to an algebraic rate, being limited by the smoothness λ of the exact solution. If proper meshes are used, that is, if the number of refinements corresponds to the polynomial degree, then any required accuracy is readily obtained.

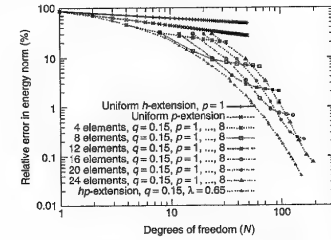


Figure 18. Comparison of the h -, p -, and hp -version: relative error in energy norm.

3.5 Model problem: The L-shaped domain

In order to illustrate the convergence characteristics of the h -, p -, and hp -extensions for category B problems, we consider an L-shaped domain in two-dimensional elasticity, under the assumption of plane strain conditions using Poisson's ratio 0.3. The reentrant edges are stress-free. In the xy coordinates system shown in Figure 12, the exact solution (up to rigid body displacement and rotation terms) corresponding to the first term of the asymptotic expansion is

$$u_x = \frac{A_1}{2G} r^{\lambda_1} [(\kappa - Q_1(\lambda_1 + 1)) \cos \lambda_1 \theta - \lambda_1 \cos(\lambda_1 - 2)\theta] \quad (42)$$

$$u_y = \frac{A_1}{2G} r^{\lambda_1} [(\kappa + Q_1(\lambda_1 + 1)) \sin \lambda_1 \theta + \lambda_1 \sin(\lambda_1 - 2)\theta] \quad (43)$$

where G is the shear modulus, $\lambda_1 = 0.544483737$, $Q_1 = 0.543075579$, and $\kappa = 1.8$. The coefficient A_1 is called a *generalized stress intensity factor*. Details are available in Szabó and Babuška (1991). The corresponding stress components are

$$\sigma_x = A_1 \lambda_1 r^{\lambda_1-1} [(2 - Q_1(\lambda_1 + 1)) \cos(\lambda_1 - 1)\theta - (\lambda_1 - 1) \cos(\lambda_1 - 3)\theta] \quad (44)$$

$$\sigma_y = A_1 \lambda_1 r^{\lambda_1-1} [(2 + Q_1(\lambda_1 + 1)) \cos(\lambda_1 - 1)\theta + (\lambda_1 - 1) \cos(\lambda_1 - 3)\theta] \quad (45)$$

$$\tau_{xy} = A_1 \lambda_1 r^{\lambda_1-1} [(\lambda_1 - 1) \sin(\lambda_1 - 3)\theta + Q_1(\lambda_1 + 1) \sin(\lambda_1 - 1)\theta] \quad (46)$$

This model problem is representative of an important class of problems. The reentrant edges are stress-free, the other boundaries are loaded by the tractions that correspond to the exact stress distribution given by equations (44) to (46). Since the exact solution is known, it is possible to compute the exact value of the potential energy from the definition of $\Pi(u_{EX})$ given by equation (3) and using $B(u_{EX}, u_{EX}) = F(u_{EX})$ from equation (1):

$$\begin{aligned} \Pi(u_{EX}) &= -\frac{1}{2} \oint [u_x(\sigma_x n_x + \tau_{xy} n_y) + u_y(\tau_{xy} n_x + \sigma_y n_y)] ds \\ &= -4.15454423 \frac{A_1 a^{2\lambda_1}}{E} \end{aligned} \quad (47)$$

where n_x, n_y are the components of the unit normal to the boundary and a is the dimension shown in the inset in Figure 12. The convergence paths for h - and p -extensions are shown in Figure 19.

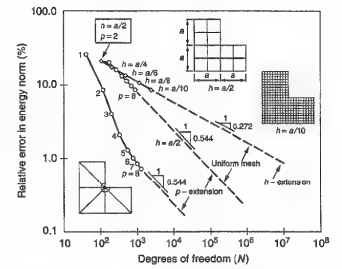


Figure 19. Convergence paths for the L-shaped domain. (From *Finite-Element Analysis*; B. Szabó and I. Babuška; Copyright (1991) John Wiley & Sons, Inc. This material is used by permission of John Wiley & Sons, Inc.)

It is seen that the asymptotic rates of convergence are exactly as predicted by the estimate (28). However, when p -extension is used on a geometric mesh, the preasymptotic rate is exponential or nearly so. This can be explained by observing that the geometric mesh shown in Figure 12 is overrefined for low polynomial degrees, hence the dominant source of the error is that part of the domain where the exact solution is smooth and hence the rate of convergence is exponential, as predicted by the estimate (29). Convergence slows to the algebraic rate for small errors, where the dominant source of error is the immediate vicinity of the singular point.

The error estimate frequently used in conjunction with p -extensions is based on the equation (6) and the use of Richardson extrapolation utilizing the a priori estimate (28). When hp -adaptivity has to be considered, local-based error estimators have to be applied; see, for example, Ainsworth and Oden (2000) and Melenk and Wohlmuth (2001). By definition, the effectivity index θ is the estimated error divided by the true error. The estimated and true errors and the effectivity indices are shown in Table 3. The parameter β is the same as that in equation (28).

4 PERFORMANCE CHARACTERISTICS

We have seen in Figure 19 that for a fixed accuracy (say 1%) there is a very substantial reduction in the number of degrees of freedom when p -extension is performed on properly designed meshes. From a practical point of view, the important consideration is the cost of computational

Table 3. L-shaped domain. Geometric mesh, 18 elements, trunk space. Plane strain, $\nu = 0.3$. Estimated and true relative errors in energy norm and effectivity index θ .

| p | N | $\Pi(u_{FE})E$ $A_1^2 a^{2p+1} t_z$ | β | | $(\epsilon_r)_E$ (%) | | θ |
|----------|----------|--|---------|------|----------------------|-------|----------|
| | | | Est.'d | True | Est.'d | True | |
| 1 | 41 | -3.886332 | — | — | 25.42 | 25.41 | 1.00 |
| 2 | 119 | -4.124867 | 1.03 | 1.03 | 8.44 | 8.46 | 1.00 |
| 3 | 209 | -4.148121 | 1.37 | 1.36 | 3.91 | 3.93 | 0.99 |
| 4 | 335 | -4.152651 | 1.33 | 1.30 | 2.09 | 2.14 | 0.98 |
| 5 | 497 | -4.153636 | 0.99 | 0.94 | 1.42 | 1.48 | 0.96 |
| 6 | 695 | -4.153975 | 0.78 | 0.68 | 1.09 | 1.17 | 0.93 |
| 7 | 929 | -4.154139 | 0.69 | 0.60 | 0.89 | 0.99 | 0.89 |
| 8 | 1199 | -4.154238 | 0.69 | 0.56 | 0.75 | 0.86 | 0.87 |
| ∞ | ∞ | -4.154470 | 0.54 | — | 0 | — | — |

resources rather than the number of degrees of freedom. The proper basis for comparing the performance characteristics of various implementations of the h - and p -versions of the finite-element method is the cost of computation. The cost has to be evaluated with respect to representative model problems, such as the L-shaped domain problem discussed in Section 3.5, given specific goals of computation, the required accuracies, and the requirement that a reasonably close estimate of the accuracy of the computed data of interest be provided. It is essential that comparisons of performance include a verification process, that is, a process by which it is ascertained that the relative errors in the data of interest are within prespecified error tolerances. Verification is understood in relation to the exact solution of the mathematical model, not in relation to some physical reality that the model is supposed to represent. The consequences of wrong engineering decisions based on erroneous information usually far outweigh the costs of verification.

Comparative performance characteristics of the h - and p -versions were first addressed in Babuska and Scapolla (1987) and Babuska and Elman (1989) through analyses of computational complexity and theoretical error estimates as well as computer timing of specific benchmark problems. The main conclusions are summarized as follows:

1. Only for the uncommon cases of very low accuracy requirements and very irregular exact solutions are low-order elements preferable to high-order elements. High-order elements typically require smaller computational effort for the same level of accuracy.
2. High-order elements are more robust than low-order elements. This point is discussed further in Section 4.1 below.
3. The most effective error control procedures combine proper mesh design coupled with progressive increase in p . For details, we refer to Rank and Babuska (1987), Babuska and Suri (1990), and Rank (1992).

4. Accuracies normally required in engineering computation can be achieved with elements of degree 8 or less for most practical problems.
5. Computation of a sequence of solutions corresponding to a hierarchic sequence of finite element spaces $S_1 \subset S_2 \subset \dots$ provides for simple and effective estimation and control of error for all data of interest, based on various types of extrapolation and extraction procedures; see, for example, Szabó and Babuska (1988), Szabó (1990), and Yosibash and Szabó (1994).

As a general rule, for problems in categories A and B (defined in Section 3.1), which include the vast majority of practical problems in solid mechanics, p extension on properly designed meshes is the most efficient general solution strategy. The performance of p -extensions in solving problems on category C is discussed in Section 5.1.1.

In the p -version, the elemental matrices are large and their computation is time-consuming. On the other hand, these operations lend themselves to parallel computation; see, for example, Rank *et al.* (2001). Furthermore, it has been shown that a substantial reduction in time can be achieved if special integration techniques are used (see Nübel, Düster and Rank, 2001), or if the hierarchic structure is sacrificed (see Melenk, Gerdes and Schwab, 2001).

4.1 Robustness

A numerical method is said to be robust when it performs well for a broad class of admissible data. For example, in the displacement formulation of linear elasticity, letting Poisson's ratio ν approach $1/2$ causes the volumetric strain ($\text{div } u$) to approach zero. This introduces constraints among the variables, effectively reducing the number of degrees of freedom, and hence causing the rate of convergence in energy norm to decrease, in some cases, very substantially. This phenomenon is called *locking*. Locking also causes problems in the recovery of the first stress invariant from the finite element solution. A similar situation exists when the thickness approaches zero in plate models based on the Reissner formulation. For a precise definition of robustness, we refer to Babuska and Suri (1992). It was shown in Vogelius (1983) that the rate of convergence in p -extensions is not influenced by $\nu \rightarrow 1/2$ on straight sided triangles. It is also known that the h -version using straight triangles does not exhibit locking, provided that $p \geq 4$. For curvilinear elements, the rate of p -convergence is slower, and for the h -version the locking problem is generally much more severe. Although the p -version is affected by membrane locking, in the range of typical plate and shell thicknesses that occur in practical engineering problems, locking effects are generally not substantial. For

an investigation of membrane locking in cylindrical shells, we refer to Pitkäranta (1992).

4.2 Example

The following example is representative of shell intersection problems. Specifically, the intersection of two cylindrical shells is considered. Referring to Figure 20, the outside radius of shell A is $R_A = 140$ mm, the outside radius of shell B is $R_B = 70$ mm. The wall thickness of shell A (resp. shell B) is $t_A = 8.5$ mm; (resp. $t_B = 7.5$ mm). The axes of the shells intersect at $\alpha = 65^\circ$. The length of shell A is 800 mm, the length of shell B, measured from the point of intersection of the axes of the shells, is 300 mm. The modulus of elasticity is $E = 72.4$ MPa, Poisson's ratio is $\nu = 0.3$.

The intersection of the outer surfaces of the shells is filleted by a 'rolling ball fillet', that is, the fillet surface is generated as if rolling a sphere of radius $r_f = 10.0$ mm along the intersection line. The mesh consists of 34 hexahedral elements. The shell intersection region, comprised of 16 elements, is the darker region shown in Figure 20. The complement is the shell region. Quasi-regional mapping utilizing 6×6 collocation points per curved face was employed.

The inside surface is loaded by a pressure \bar{p} . In order to establish equilibrium, a normal traction T_n is applied on the surface S_B , which is the surface of intersection between shell B and a plane perpendicular to its axis:

$$T_n = \frac{\bar{p}(R_B - t_B)^2}{t_B(2R_B - t_B)} \quad (48)$$

The yz plane is a plane of symmetry. The other surfaces are traction-free. Appropriate rigid body constraints were imposed in order to prevent motion parallel to the plane of symmetry.

The objective is to estimate the magnitude of the maximal von Mises stress to within 5% relative error. In the

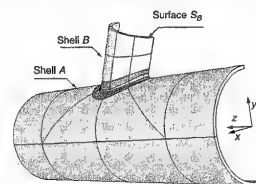


Figure 20. Example: Shell intersection problem. The darker region, comprised of 16 elements, is the shell intersection region.

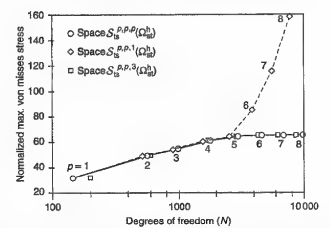


Figure 21. Example: Shell intersection problem. Convergence of the maximum von Mises stress normalized with respect to the applied pressure \bar{p} . The estimated limit is 64.7. The maximum occurs in the shell intersection region.

shell intersection region, the solution varies substantially over distances comparable to the thickness. Therefore, dimensional reduction cannot be justified for this region. Fully three-dimensional elements, that is, elements based on the trunk spaces $S_0^{p,p,p,1}(\Omega_h^0)$ with $p = 1, 2, \dots, 8$ were used in the shell intersection region, whereas the anisotropic spaces $S_0^{p,p,p,1}(\Omega_h^0)$ were used in the shell region. The computations were performed with StressCheck [2].

The results are shown in Figure 21. Very strong convergence to the estimated limit value of 64.7 is observed when the isotropic spaces $S_0^{p,p,p}(\Omega_h^0)$ are employed. This is true also for the anisotropic spaces $S_0^{p,p,p,1}(\Omega_h^0)$ for $q \geq 2$ but not for $q = 1$. The reason is that $q = 1$ implies kinematic assumptions similar to those of the Naghdi shell theory. This introduces an artificial singularity along the faces where q changes abruptly from 8 to 1. Essentially, this is a modeling error in the sense it pertains to the question of whether and where a particular shell model is applicable, given that the goal is to approximate some functional of the exact solution of the underlying fully three-dimensional problem. Some aspects of this problem have been addressed in Schwab (1996) and Actis, Szabó and Schwab (1999). This example illustrates the importance of convergence tests on the data of interest, including tests on the choice of dimensionally reduced models.

5 APPLICATIONS TO NONLINEAR PROBLEMS

5.1 Elastoplasticity

The p - and hp -versions of the finite element method have been widely accepted as efficient, accurate, and flexible

methods for analyzing linear problems in computational mechanics. On the other hand, applications of the p - and hp -versions to nonlinear problems are relatively recent and hence less well known. Considering for instance, the J_2 flow theory of elastoplasticity, a loss of regularity is expected along the boundary of the plastic zone. Following the classification of Section 3.1, this problem is of Class C, that is, it has an unknown line (in 2D) or surface (in 3D) of singular behavior in the interior of the domain. Therefore, only an algebraic rate of convergence can be expected. However, this asymptotic rate does not give information on the preasymptotic behavior, that is, on the accuracy of a p -extension for a finite number of degrees of freedom, and especially on the question of computational investment for a desired accuracy of quantities of engineering interest.

To shed some light on this question, we will investigate the deformation theory of plasticity, first proposed by Hencky (1924), as a very simple model problem for elastoplasticity. For a detailed description and numerical investigation of this model problem, see Szabó, Actis and Holzer (1995) and Düster and Rank (2001). We refer to Holzer and Yosibash (1996), Düster and Rank (2002), and Düster *et al.* (2002) for a study of the more complex and physically more realistic flow theory of plasticity, where each load integration step in an incremental analysis can be considered equivalent to the model problem investigated in the following section.

5.1.1 A benchmark problem

As a numerical example, we again use the structure of Figure 9 in Section 2.4.2 showing a quarter of a square plate with central hole and unit thickness, loaded now by a uniform tension of magnitude $T_x = 450$ MPa. The dimensions of the plate are chosen to be $b = h = 10$ mm and the radius is set to $R = 1$ mm. The material is now assumed to be elastic, perfectly plastic and plane strain conditions are assumed. The shear modulus is $\mu = 80193.8$ MPa, the bulk modulus is $\kappa = 164206.0$ MPa, and the yield stress is $\sigma_0 = 450.0$ MPa. This problem was defined by Stein (2002) as a benchmark problem for the German research project 'Adaptive finite-element methods in applied mechanics'.

To find an approximate solution for the given benchmark, we use the p -version based on the tensor product space $S_{p,p}^0(\Omega_h^0)$ taking advantage of the blending function method. Three different meshes with 2, 4 and 10 p -elements have been chosen (see Figure 22). A series of computations for polynomial degrees $p \leq 17$ for the mesh with 2 elements and $p \leq 9$ for the meshes with 4 and 10 elements was performed. In order to make a comparison with an adaptive h -version, we refer to the results of Barthold,



Figure 22. Three meshes with 2, 4, and 10 p -elements. (Reprinted from *Comput. Methods Appl. Mech. Engng.*, 190, A. Düster and E. Rank, The p -version of the finite-element method is compared to an adaptive h -version for the deformation theory of plasticity, 1925–1935, Copyright (2001), with permission from Elsevier.)

Schmidt and Stein (1997, 1998) and Stein *et al.* (1997). The computations there were performed with the Q1-P0 element differing from the well known bilinear quadrilateral element by including an additional, elementwise constant pressure degree of freedom. A mesh consisting of 64 Q1-P0 elements was refined in 10 steps using the equilibrium criterion, yielding 875 elements with 1816 degrees of freedom (see Figure 23). In Barthold, Schmidt and Stein (1997, 1998) and Stein *et al.* (1997), the results of a sequence of graded meshes and a reference solution obtained with 24200 Q1-P0 elements with a corresponding number of 49062 degrees of freedom are also given. Comparing the results of the uniform p -version with those of the h -version based on a sequence of graded meshes, we observe that the efficiency of the p -version is superior (see Figures 24, 25). The discretization with 4 elements, $p = 9$, and 684 degrees of freedom provides an accuracy that cannot be reached by the h -version, even when using 4096 Q1-P0 elements with 8320 degrees of freedom. Even compared to an h -refinement, resulting in an adapted mesh with 875 Q1-P0 elements, it can be seen that a uniform p -version is much more accurate. Although the p -version is significantly more elaborate than the h -version, when comparing the computational effort per degree of freedom, investigations on the computational cost to obtain highly accurate results have clearly shown a superiority of high-order elements. For further information, including three-dimensional examples of the J_2 flow theory with nonlinear isotropic hardening, we refer to Düster and Rank (2001, 2002), Düster (2002), Düster *et al.* (2002), Rank *et al.* (2002), and Stein (2002).

5.1.2 An industrial application

The following example is concerned with a structural component of moderate complexity, called a *dragbrace fitting*, shown in Figure 26. This part is representative of structural components used in the aerospace sector in that relatively thin plate-like regions are reinforced by integrally machined stiffeners. The overall dimensions are length $L = 219.6$ mm and width $w = 115$ mm. The material is typically aluminum or titanium, which exhibit strain hardening. For

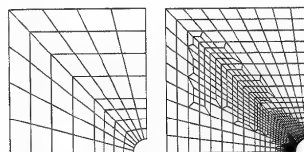


Figure 23. Initial mesh with 64 Q1-P0 elements and adapted mesh with 875 Q1-P0 elements (see Barthold, Schmidt and Stein, 1997). (Reprinted from *Comput. Methods Appl. Mech. Engng.*, 190, A. Düster and E. Rank, The p -version of the finite-element method compared to an adaptive h -version for the deformation theory of plasticity, 1925–1935, Copyright (2001), with permission from Elsevier.)

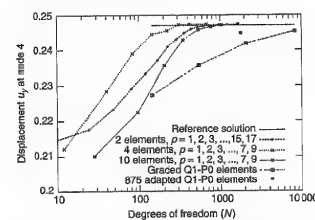


Figure 24. Displacement u_x at node 4. (Reprinted from *Comput. Methods Appl. Mech. Engng.*, 190, A. Düster and E. Rank, The p -version of the finite-element method compared to an adaptive h -version for the deformation theory of plasticity, 1925–1935, Copyright (2001), with permission from Elsevier.)

the purposes of this example, an elastic-perfectly plastic material was chosen because it poses a more challenging problem from the numerical point of view. The material properties are those of an ASTM A-36 steel; the yield point is 248 MPa, the modulus of elasticity is $E = 200$ GPa, and Poisson's ratio is $\nu = 0.295$. The mathematical model is based on the deformation theory of plasticity.

The lugs A and B are fully constrained and sinusoidally distributed normal tractions are applied through lugs C and D. The resultants of the tractions are F and $2F$ respectively, acting in the negative x direction as the dark region shown schematically in Figure 26. The goal of the computation is to determine the extent of the plastic zone, given that $F = 5.5$ kN. The mesh consists of 2 tetrahedral elements, 22 pentahedral elements, and 182 hexahedral elements.

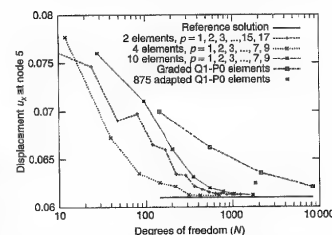


Figure 25. Displacement u_x at node 5. (Reprinted from *Comput. Methods Appl. Mech. Engng.*, 190, A. Düster and E. Rank, The p -version of the finite-element method compared to an adaptive h -version for the deformation theory of plasticity, 1925–1935, Copyright (2001), with permission from Elsevier.)

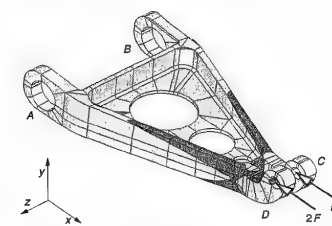


Figure 26. Example: Dragbrace fitting. Elastic-plastic solution, $p = 7$, trunk space, $N = 49894$. In the dark region, the equivalent strain exceeds the yield strain.

The region of primary interest is the neighborhood of the loaded lugs. The results of linear analysis indicate that the maximal von Mises stress in this region is 1040 MPa, that is, 4.2 times the yield stress. Therefore, nonlinear analysis has to be performed. The region where the equivalent strain exceeds the yield strain is shown in Figure 26. The computations were performed with StressCheck.

5.2 Geometric nonlinearity

The following example illustrates an application of the p -version to a geometrically nonlinear problem. In geometrically nonlinear problems, equilibrium is satisfied in the

deformed configuration. The constitutive laws establish a relationship either between the Piola-Kirchhoff stress tensor and the Euler-Lagrange strain tensor or the Cauchy stress tensor and the Almansi strain tensor. The formulation in this example is based on the Cauchy stress and the Almansi strain; see Noel and Szabó (1997). The mapping functions given by equation (25) are updated iteratively by the displacement vector components. For example, at the i th iteration, the x -coordinate is mapped by

$$x^{(i)} = Q_1^i(\xi, \eta, \zeta) + u_x^{(i)}(\xi, \eta, \zeta) \quad (49)$$

It is known that when a thin elastic strip is subjected to pure bending, it deforms so that the curvature is constant and proportional to the bending moment:

$$\frac{1}{\rho} = \frac{M}{EI} \quad (50)$$

where ρ is the radius of curvature, M is the bending moment, E is the modulus of elasticity, and I is the moment of inertia. Poisson's ratio ν is zero. In this example, a thin strip of length $L = 100$ mm, thickness $t = 0.5$ mm, and width $b = 5$ mm is subjected to normal tractions on Face A shown in Figure 27, which correspond to M chosen so that $\rho = L/2\pi$:

$$T_n = -\frac{2\pi E}{L} \bar{y} \quad (51)$$

where \bar{y} is measured from the mid surface in the direction of the normal in the current configuration. The three displacement vector components are set to zero on Face B. Three hexahedral elements were used. The anisotropic space $S_0^{p,p,1}(\Omega_0^h)$ described in Section 2.3 was used. Mapping was by the blending function method using 6×6 collocation points in the quasi-regional mapping procedure described by Kirdályfalvi and Szabó (1997). The computations were performed with StressCheck. The load T_n was applied in 20 equal increments. The final deformed configuration, a nearly perfect cylindrical body, is shown in Figure 27. The exact solution of a perfectly cylindrical middle surface (the elastica) is the limiting case with respect to the thickness approaching zero.

This example illustrates the following: (a) In the p -version, very large aspect ratios can be used. (b) Quasi-regional mapping, which is an extension of isoparametric mapping combined with the blending function method, is capable of providing a highly accurate representation of the geometrical description with very few elements over large deformations. In this example, Face A was rotated 360 degrees relative to its reference position.

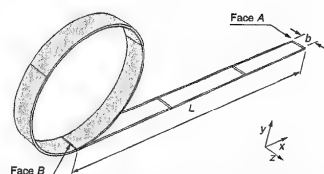


Figure 27. Example: Thin elastic strip. Geometrically nonlinear solution. Three hexahedral elements, anisotropic space $S_0^{p,p,1}(\Omega_0^h)$, $N = 684$.

6 OUTLOOK

Although implementations of the p -version are available in a number of commercial finite element computer codes, widespread applications of the p -version in professional practice have been limited by three factors:

1. The infrastructure of the most widely used FEA software products was designed for the h -version, and cannot be readily adapted to meet the technical requirements of the p -version.
2. In typical industrial problems, finite element meshes are generated by automatic mesh generators that produce very large numbers of tetrahedral elements mapped by low-order (linear or quadratic) polynomial mapping functions. When the mapping functions are of low degree, the use of high-order elements is generally not justified. This point was illustrated in Section 2.4.2. Nevertheless, numerous computational experiments have shown that p -extension performed on tetrahedral meshes up to $p = 4$ or $p = 5$ provides efficient means of verification for the computed data when the mappings are proper, that is, the Jacobian determinant is positive over every element. Experience has shown that many commercial mesh generators produce improperly mapped elements. As mesh generators improve and produce fewer elements and more accurate mappings, this obstacle will be gradually removed.
3. The demand for verified information in industrial applications of FEMs has been generally weak; however, as computed information is becoming an increasingly important part of the engineering decision-making process, the demand for verified data, and hence the importance of the p -version, is likely to increase.

At present, the p -version is employed in industrial applications mainly where it provides unique technical capabilities. Some examples are: (a) Analysis of mechanical and structural components comprised of plate- and shell-like regions where dimensional reduction is applicable, and solid regions where fully three-dimensional representation is necessary. An example of this kind of domain is shown in Figure 26 where it would not be feasible to employ fully automatic mesh generators because the fillets would cause the creation of an excessive number of tetrahedral elements. On the other hand, if the fillets were omitted, then the stresses could not be determined in the most critical regions. (b) Two- and three-dimensional linear elastic fracture mechanics where p -extensions on geometric meshes, combined with advanced extraction procedures, provide verified data very efficiently; see, for example, Szabó and Babuska (1988) and Andersson, Falk and Babuska (1990). (c) Plate and shell models where the robustness of the p -version and its ability to resolve boundary layer effects are important; see, for example, Babuska, Szabó and Actis (1992), Actis, Szabó and Schwab (1999), and Rank, Krause and Preusch (1998). (d) Analysis of structural components made of composite materials where special care must be exercised in choosing the mathematical model; large aspect ratios must be used and geometric as well as material nonlinear effects may have to be considered; see Engelstad and Actis (2003). (e) Interpretation of experimental data where strict control of the errors of discretization (as well as the experimental errors) is essential for proper interpretation of the results of physical experiments.

The p -version continues to be a subject of research aimed at broadening its application to new areas. Only a few of the many important recent and ongoing research activities can be mentioned here. Application of the p - and hp -versions to mechanical contact is discussed in Páczelt and Szabó (2002) and the references listed therein. The problem of hp -adaptivity was addressed in the papers Demkowicz, Oden and Rachowicz (1989), Oden *et al.* (1989), Rachowicz, Oden and Demkowicz (1989), and Demkowicz, Rachowicz and Devloo (2002). The design of p -adaptive methods for elliptic problems was addressed in Bertóti and Szabó (1998). The problem of combining p - and hp -methods with boundary element methods (BEMs) for the solution of elastic scattering problems is discussed in Demkowicz and Oden (1996). Further information on coupling of FEM and BEM can be found in this encyclopedia (see Chapter 13, this Volume). Application of hp -adaptive methods to Maxwell equations was reported in Rachowicz and Demkowicz (2002).

ACKNOWLEDGMENTS

The writers wish to thank Dr Ricardo Actis of Engineering Software Research and Development, Inc., St. Louis Missouri, USA for assistance provided in connection with the examples computed with StressCheck and Professor István Páczelt of the University of Miskolc, Hungary, for helpful comments on the manuscript.

NOTES

- [1] Waterloo Maple Inc., 57 Erb Street West, Waterloo, Ontario, Canada (www.maplesoft.com). The worksheet can be obtained from the Lehrstuhl für Bauinformatik, Technische Universität München, Germany (www.inf.bv.tum.de/~duester).
- [2] StressCheck is a trademark of Engineering Software Research and Development, Inc., St. Louis, Missouri, USA (www.esrd.com).

REFERENCES

- Actis Szabó BA and Schwab C. Hierarchic models for laminated plates and shells. *Comput. Methods Appl. Mech. Eng.* 1999; 172:79–107.
- Ainsworth M and Oden JT. *A Posteriori Error Estimation in Finite Element Analysis*. John Wiley & Sons: New York, 2000.
- Andersson B, Falk U and Babuska I. *Accurate and Reliable Determination of Edge and Vertex Stress Intensity Factors*. Report FFA TN 1990-28, The Aeronautical Research Institute of Sweden: Stockholm, 1990.
- Babuska I and Elman HC. Some aspects of parallel implementation of the finite element method on message passing architectures. *J. Comput. Appl. Math.* 1989; 27:157–189.
- Babuska I and Scapolla T. Computational aspects of the h -, p - and hp -versions of the finite element method. In *Advances in Computer Methods in Partial Differential Equations – VI*, Voinovetsky R and Stepleman RS (eds), International Association for Mathematics and Computer Simulation (IMACS), 1987; 233–240.
- Babuska I and Strouboulis T. *The Finite Element Method and its Reliability*. Oxford University Press: Oxford, 2001.
- Babuska I and Suri M. The p - and hp -versions of the finite element method, an overview. *Comput. Methods Appl. Mech. Eng.* 1990; 80:5–26.
- Babuska I and Suri M. On locking and robustness in the finite element method. *SIAM J. Numer. Anal.* 1992; 29:1261–1293.
- Babuska I, Szabó BA and Actis RL. Hierarchic models for laminated composites. *Int. J. Numer. Methods Eng.* 1992; 33:503–535.
- Barthold FF, Schmidt M and Stein E. Error estimation and mesh adaptivity for elastoplastic deformations. In *Proceedings of*

- the 5th International Conference on Computational Plasticity, Compas V, Barcelona, 1997.
- Barthold FJ, Schmidt M and Stein E. Error indicators and mesh refinements for finite-element-computations of elastoplastic deformations. *Comput. Mech.* 1998; 22:225–238.
- Bertóti E and Szabó B. Adaptive selection of polynomial degrees on a finite element mesh. *Int. J. Numer. Methods Eng.* 1998; 42:561–578.
- Bröker H. *Integration von geometrischer Modellierung und Berechnung nach der p-Version der FEM*. PhD thesis, Lehrstuhl für Bauinformatik, Technische Universität München, 2001; published in Shaker Verlag: Aachen, ISBN 3-8265-9653-6, 2002.
- Chen Q and Babuška I. Approximate optimal points for polynomial interpolation of real functions in an interval and in a triangle. *Comput. Methods Appl. Mech. Eng.* 1995; 128:405–417.
- Chen Q and Babuška I. The optimal symmetrical points for polynomial interpolation of real functions in the tetrahedron. *Comput. Methods Appl. Mech. Eng.* 1996; 137:89–94.
- Demkowicz LF and Oden JT. Application of hp-adaptive BE/FE methods to elastic scattering. *Comput. Methods Appl. Mech. Eng.* 1996; 133:287–318.
- Demkowicz LF, Oden JT and Rachowicz W. Toward a universal h-p adaptive finite element strategy, Part 1. Constrained approximation and data structure. *Comput. Methods Appl. Mech. Eng.* 1989; 77:79–112.
- Demkowicz LF, Rachowicz W and Devloo PH. A fully automatic hp-adaptivity. *J. Sci. Comput.* 2002; 17:127–155.
- Düster A. *High Order Finite Elements for Three-Dimensional, Thin-Walled Nonlinear Continua*. PhD thesis, Lehrstuhl für Bauinformatik, Technische Universität München, 2001; published in Shaker Verlag: Aachen, ISBN 3-8322-0189-0, 2002.
- Düster A, Bröker H and Rank E. The p-version of the finite element method for three-dimensional curved thin walled structures. *Int. J. Numer. Methods Eng.* 2001; 52:673–703.
- Düster A, Niggel A, Nübel V and Rank E. A numerical investigation of high-order finite elements for problems of elastoplasticity. *J. Sci. Comput.* 2002; 17:397–404.
- Düster A and Rank E. The p-version of the finite element method compared to an adaptive h-version for the deformation theory of plasticity. *Comput. Methods Appl. Mech. Eng.* 2001; 190:1925–1935.
- Düster A and Rank E. A p-version finite element approach for two- and three-dimensional problems of the J_2 flow theory with non-linear isotropic hardening. *Int. J. Numer. Methods Eng.* 2002; 53:49–63.
- Engelstad SP and Actis RL. Development of p-version handbook solutions for analysis of composite bonded joints. *Comput. Math. Appl.* 2003; 46:81–94.
- Gordon WJ and Hall ChA. Construction of curvilinear co-ordinate systems and applications to mesh generation. *Int. J. Numer. Methods Eng.* 1973a; 7:461–477.
- Gordon WJ and Hall ChA. Transfinite element methods: blending function interpolation over arbitrary curved element domains. *Numer. Math.* 1973b; 21:109–129.
- Grisvard P. *Elliptic Problems in Nonsmooth Domains*. Pitman Advanced Pub. Program Boston, 1985.
- Hencky H. Zur Theorie plastischer Deformationen und der hierdurch im Material hervorgerufenen Nebenspannungen. In *Proceedings of the 1st International Congress on Applied Mechanics*, Delft, 1924.
- Holzer S and Yosibash Z. The p-version of the finite element method in incremental elasto-plastic analysis. *Int. J. Numer. Methods Eng.* 1996; 39:1859–1878.
- Királyfalvi G and Szabó BA. Quasi-regional mapping for the p-version of the finite element method. *Finite Elem. Anal. Des.* 1997; 27:85–97.
- Melenk M, Gerdas K and Schwab C. Fully discrete hp-finite elements: fast quadrature. *Comput. Methods Appl. Mech. Eng.* 2001; 190:4339–4369.
- Melenk M and Wohlmuth B. On residual-based a-posteriori error estimation in hp-FEM. *Adv. Comput. Math.* 2001; 15:311–331.
- Noel AT and Szabó BA. Formulation of geometrically non-linear problems in the spatial reference frame. *Int. J. Numer. Methods Eng.* 1997; 40:1263–1280.
- Nübel V, Düster A and Rank E. Adaptive vector integration as an efficient quadrature scheme for p-version finite element matrices. In *Proceedings of the European Conference on Computational Mechanics*, Cracow, 2001.
- Oden JT, Demkowicz LF, Rachowicz W and Westermann T. Toward a universal h-p adaptive finite element strategy, Part 2. A posteriori error estimation. *Comput. Methods Appl. Mech. Eng.* 1989; 77:113–180.
- Páczelt I and Szabó B. Solution of contact optimization problems of cylindrical bodies using hp-FEM. *Int. J. Numer. Methods Eng.* 2002; 53:123–146.
- Pitkäranta J. The problem of membrane locking in finite element analysis of cylindrical shells. *Numer. Math.* 1992; 61:523–542.
- Rachowicz W, Oden JT and Demkowicz LF. Toward a universal h-p adaptive finite element strategy, Part 3. Design of h-p meshes. *Comput. Methods Appl. Mech. Eng.* 1989; 77:181–212.
- Rachowicz W and Demkowicz LF. An hp-adaptive finite element method for electromagnetics – Part II: A 3D implementation. *Int. J. Numer. Methods Eng.* 2002; 53:147–180.
- Rank E. Adaptive remeshing and h-p domain decomposition. *Comput. Methods Appl. Mech. Eng.* 1992; 101:299–313.
- Rank E and Babuška I. An expert system for the optimal mesh design in the hp-version of the finite element method. *Int. J. Numer. Methods Eng.* 1987; 24:2087–2106.
- Rank E, Bröker H, Düster A, Krause R and Rücker M. The p-version of the finite element method for structural problems. In *Error-Controlled Adaptive Finite Elements in Solid Mechanics*, Stein E (ed.). John Wiley & Sons: New York, 2002; 263–307.
- Rank E, Krause R and Preuss K. On the accuracy of p-version elements for the Reissner-Mindlin plate problem. *Int. J. Numer. Methods Eng.* 1998; 43:51–67.
- Rank E, Rücker M, Düster A and Bröker H. The efficiency of the p-version finite element method in a distributed computing environment. *Int. J. Numer. Methods Eng.* 2001; 52:589–604.
- Schwab Ch. *p- and hp-Finite Element Methods*. Oxford University Press: Oxford, 1998.
- Schwab C. A-posteriori modeling error estimation for hierarchic plate models. *Numer. Math.* 1996; 74:221–259.
- Stein E. *Error-Controlled Adaptive Finite Elements in Solid Mechanics*. John Wiley & Sons, 2002.
- Stein E, Barthold FJ, Ohnimas S and Schmidt M. Adaptive finite elements in elastoplasticity with mechanical error indicators and neumann-type estimators. In *Proceedings of the Workshop on New Advances in Adaptive Computational Mechanics*, Cachan, 1997.
- Szabó BA, Actis R and Holzer S. Solution of elastic-plastic stress analysis problems by the p-version of the finite element method. In *Modeling, Mesh Generation, and Adaptive Numerical Methods for Partial Differential Equations*, IMA Volumes in Mathematics and its Applications, vol. 75, Babuška I and Flaherty J (eds). Springer, 1995; 395–416.
- Szabó BA and Babuška I. Computation of the amplitude of stress singular terms for cracks and reentrant corners. In *Fracture Mechanics: Nineteenth Symposium*, Cruse T (ed.). ASTM STP 969. Philadelphia, 1988; 101–124.
- Szabó BA and Babuška I. *Finite Element Analysis*. John Wiley & Sons: New York, 1991.
- Szabó BA. The p- and hp-versions of the finite element method in solid mechanics. *Comput. Methods Appl. Mech. Eng.* 1990; 80:185–195.
- Vogelius M. An analysis of the p-version of the finite element method for nearly incompressible materials – uniformly valid optimal estimates. *Numer. Math.* 1985; 41:39–53.
- Yosibash Z and Szabó B. Convergence of stress maxima in finite element computations. *Commun. Numer. Methods Eng.* 1994; 10:683–697.

FURTHER READING

Simo JC and Hughes TJR. *Computational Inelasticity*. Springer-Verlag: New York, 1998.

Chapter 6 Spectral Methods

Claudio Canuto¹ and Alfio Quarteroni²

¹ Politecnico di Torino, Turin, Italy

² MOX, Politecnico di Milano, Milan, Italy and IACS, School of Mathematics, EPFL, Lausanne, Switzerland

| | |
|---|-----|
| 1 Introduction | 141 |
| 2 Fourier Methods | 141 |
| 3 Algebraic Polynomial Expansion | 143 |
| 4 Algebraic Expansions on Triangles | 145 |
| 5 Stokes and Navier–Stokes Equations | 146 |
| 6 Advection Equations and Conservation Laws | 148 |
| 7 The Spectral Element Method | 150 |
| 8 The Mortar Method | 152 |
| References | 154 |

1 INTRODUCTION

In the past three decades, spectral methods have evolved from their noble ancestor, the Fourier method based on trigonometric expansions, through the more flexible Galerkin method with Gaussian integration, all the way maintaining their most distinguished feature: the very high rate of convergence.

They are numerical methods for solving boundary-value problems for partial differential equations.

For the reader's convenience, we will gradually approach this subject by first addressing the case of periodic problems, where the so-called Fourier methods are used. Then we turn to nonperiodic problems and address collocation approximations based on algebraic polynomial expansions. The different concepts are first explained on one-dimensional intervals. Then we address the case of a square

or a cube or a simplex, and finally the case of more complex geometrical domains. We illustrate the case of elliptic equations, Stokes and Navier–Stokes equations, and then advection equations and conservation laws.

2 FOURIER METHODS

In their early stage, spectral methods were designed to approximate the periodic solution of partial differential equations by a truncated Fourier series. If

$$u(x) = \sum_{k=-\infty}^{+\infty} u_k \varphi_k(x), \quad \varphi_k(x) = \frac{1}{\sqrt{2\pi}} e^{ikx}$$

is the unknown solution, the numerical solution is sought in the form

$$u_N(x) = \sum_{k=-N/2}^{(N/2)-1} u_{N,k} \varphi_k(x)$$

where N is an (even) integer that dictates the size of the approximate problem. Note that the unknowns are represented by the Fourier coefficients $\{u_{N,k}\}$. Potentially, this approximation has a tremendously high quality, since for all $0 \leq k \leq s$, there exists a positive constant $C_{k,s}$ such that

$$\inf_{v_N \in S_N} \|u - v_N\|_{H^s(0,2\pi)} \leq C_{k,s} N^{k-s} \|u\|_{H_s^s(0,2\pi)} \quad (1)$$

provided u belongs to the Sobolev space $H_s^s(0,2\pi)$ of periodic functions having s derivatives in $L^2(0,2\pi)$. Here, $S_N = \text{span}\{\varphi_k | -N/2 \leq k \leq (N/2) - 1\}$ denotes the space of the finite Fourier series of order N . This abstract

approximation property is reflected by a corresponding error estimate for the difference $u - u_N$. Actually, in the most classical approach, the spectral Fourier method consists of approximating a given PDE, say

$$Lu(x) = f(x), \quad x \in \Omega \quad (2)$$

with $\Omega = (0, 2\pi)$, where L is a differential operator and f is a given 2π -periodic function, by requiring the L^2 -projection of the residual upon the subspace S_N to vanish, that is,

$$\text{find } u_N \in S_N \quad \text{s.t.} \quad (Lu_N - f, \varphi) = 0 \quad \forall \varphi \in S_N \quad (3)$$

Here, $(v, w) = \int_{\Omega} v \bar{w}$ denotes the $L^2(\Omega)$ inner product. If L is a constant coefficient operator, this yields an embarrassingly simple problem. As a matter of fact, owing to the L^2 -orthogonality of the functions $\{\varphi_k\}$, that is, $(\varphi_k, \varphi_m) = \delta_{km}$, $\forall k, m \in \mathbb{Z}$, equation (3) yields, after Fourier-transforming the residual $Lu_N - f$, the following set of explicit equations for the unknowns:

$$\lambda_k u_{N,k} = \hat{f}_k, \quad -\frac{N}{2} \leq k \leq \frac{N}{2} - 1 \quad (4)$$

where $\hat{f}_k = (f, \varphi_k)$ is the k th Fourier coefficient of f , while λ_k is the k th eigenvalue of L . For instance, if

$$L = -D(\alpha D) + \beta D + \gamma I \quad (\text{with } \alpha, \beta, \gamma \text{ constants}) \quad (5)$$

where D denotes differentiation with respect to x and I the identity, then $\lambda_k = \alpha k^2 + i\beta k + \gamma$.

Moreover, in this special case, u_N indeed coincides with the truncated Fourier series of order N of the exact solution u , thus the bound (1) (with $v_N = u_N$) provides an error estimate.

However, the one that we have just described is an overly fortunate circumstance. Should indeed some of the coefficients α, β , or γ be functions of x (or, even worse, of u , yielding a nonlinear equation), then convolution sums between the unknown frequency coefficients $\{u_{N,k}\}$ and the Fourier coefficients of α, β, γ will arise, and the diagonal structure of equation (4) would be lost. A variant of the projection approach (3) can be based on evaluating the convolution sums by *discrete Fourier transform*. This requires introducing equally spaced nodes, $x_j = \pi j/N$, $j = 0, \dots, N-1$, then replacing the exact integrals in (3) by numerical integration; the resulting scheme is

$$\text{find } u_N \in S_N \quad \text{s.t.} \quad (Lu_N^c - f, \varphi)_N = 0 \quad \forall \varphi \in S_N \quad (6)$$

where $(v, w)_N = (2\pi/N) \sum_{j=0}^{N-1} v(x_j) \bar{w}(x_j)$ is the Gaussian approximation of the scalar product (v, w) . The exactness of the Gaussian approximation on S_N , namely, the

property that $(v, w)_N = (v, w)$, $\forall v, w \in S_N$, enables us to recover from (6) a *collocation* formulation $L_N u_N^c = f$ at all nodes x_j , where L_N is obtained from L by replacing each derivative by the corresponding so-called *pseudospectral derivative*. This means that for any smooth function v , Dv is replaced by $D(I_N v)$, where

$$I_N v \in S_N, \quad I_N v(x) = \sum_{k=-N/2}^{(N/2)-1} v_k^* \varphi_k(x), \quad v_k^* = (v, \varphi_k)_N \quad (7)$$

is the interpolant of v at the nodes $\{x_j\}$.

The interpolation error satisfies, for $0 \leq k \leq s$, $s \geq 1$,

$$\|v - I_N v\|_{H^s(0, 2\pi)} \leq C_s N^{k-s} \|v\|_{H_p^s(0, 2\pi)} \quad (8)$$

and so does the collocation error $u - u_N^c$. A consequence of (8) (when $k=1$) is that the error on the pseudospectral derivative $\|v' - (I_N v)'\|_{L^2(0, 2\pi)}$ decreases like a constant time N^{1-s} , provided that $v \in H_p^s(0, 2\pi)$ for some $s \geq 1$. Indeed, one can even prove that

$$\|Dv - D(I_N v)\|_{L^1(0, 2\pi)} \leq C(\eta) N e^{-N\eta/2} \quad \forall 0 < \eta < \eta_0$$

provided that v is analytic in the strip $|\text{Im } z| < \eta_0$. This exponential rate of convergence is often referred to as *spectral convergence*, as it is a distinguishing feature of spectral methods.

There is, however, a major difference between the collocation approach and the L^2 -projection approach (3). In the latter, the unknowns are the frequency coefficients $\{u_{N,k}\}$ of u_N , whereas in the collocation approach one looks for the nodal values $\{u_j = u_N^c(x_j)\}$ of u_N^c . These values may be interpreted as the coefficients of u_N^c with respect to the trigonometric *Lagrange basis* associated with the nodes x_j ; indeed, observing that $u_N^c = I_N u_N^c$, using (7) and exchanging summations over k and j , one gets

$$u_N^c(x) = \sum_{j=0}^{N-1} u_j \frac{2\pi}{N} \sum_{k=-N/2}^{(N/2)-1} \varphi_k(x_j) \varphi_k(x) = \sum_{j=0}^{N-1} u_j \psi_j(x)$$

where $\psi_j \in S_N$ satisfies $\psi_j(x_m) = \delta_{jm}$, $0 \leq j, m \leq N-1$. A modal representation is used in the former case (Fourier), whereas a nodal one is adopted in the latter (collocation).

The same approach can be pursued for boundary-value problems set on multidimensional intervals $\Omega = (0, 2\pi)^d$, $d = 2, 3$ by tensorizing basis functions and collocation nodes.

Fourier methods represent the most classical approach in spectral methods. The interested reader can find a comprehensive coverage of the subject in the monographs Gottlieb and Orszag (1977) and Canuto *et al.* (1988).

3 ALGEBRAIC POLYNOMIAL EXPANSION

When a boundary-value problem with nonperiodic data (of Dirichlet, Neumann, or mixed type) has to be solved numerically, the trigonometric expansion is no longer adequate to guarantee high order of accuracy. Then, Jacobi orthogonal polynomials are used to provide orthogonal bases for the approximation space.

The finite dimensional space \mathbb{P}_N is now made of algebraic polynomials of degrees less than or equal to N .

The historical approach, inspired by the Fourier method, aimed at expanding the approximate solution with respect to a basis of orthogonal polynomials

$$u_N(x) = \sum_{k=0}^N u_{N,k} p_k(x) \quad (9)$$

where $u_{N,k}$ now represent the unknown frequency coefficients.

The matter of choice were the Chebyshev polynomials, $p_k(x) = T_k(x) = \cos(k\theta)$, $\theta = \cos^{-1}(x)$, $-1 \leq x \leq 1$, owing to their analogy with trigonometric polynomials. Since the Chebyshev basis does not necessarily match the boundary requirement (as $T_k(1) = 1$, $T_k(-1) = (-1)^k$, $\forall k \geq 0$), one device consists of projecting the equation residual on the reduced space \mathbb{P}_{N-2} , enforcing the boundary conditions afterward. For instance, for a Dirichlet boundary-value problem like (2), where now $\Omega = (-1, 1)$, and Dirichlet boundary conditions $u(-1) = u_-$, $u(1) = u_+$, the solution (9) is required to satisfy the following equations:

$$(Lu_N - f, T_k)_\omega = 0, \quad 0 \leq k \leq N-2 \quad (10)$$

$$u_N(-1) = u_-, \quad u_N(1) = u_+$$

This modal approach was termed the *Lanczos-Tau method*. The symbol $(u, v)_\omega = \int_{-1}^1 uv \omega dx$ is the so-called *weighted scalar product* with respect to the Chebyshev weight function $\omega(x) = (1-x^2)^{-1/2}$, $-1 < x < 1$. The weighted scalar product is used, instead of the more traditional one (\cdot, \cdot) , in order to take advantage (to the highest possible extent) of the Chebyshev orthogonality,

$$(T_k, T_m)_\omega = 0 \quad \text{if } k \neq m$$

$$(T_0, T_0)_\omega = \pi$$

$$(T_k, T_k)_\omega = \frac{\pi}{2} \quad \forall k \geq 1$$

When L has constant coefficients, the Lanczos-Tau problem (10) yields an algebraic system for the frequency coefficients $\{u_{N,k}\}$ with a structured matrix for which efficient

diagonalization algorithms can be devised, a circumstance that is also featured by the multidimensional problems that are generated by tensorization.

However, this is not general enough, as this structure gets lost for a more general kind of differential operators. A more flexible approach (in analogy with what was done in the Fourier case) consists of adopting a nodal representation of u_N at selected Gauss-Lobatto nodes $x_j = \cos(\pi j/N)$, $j = 0, \dots, N$, then looking for a standard Galerkin approximation with integrals replaced by Gauss-Lobatto integration:

$$(u, v)_N = \sum_{j=0}^N \alpha_j u(x_j) v(x_j) \quad (11)$$

where $\alpha_j = (\pi/N)$ for $j = 1, \dots, N-1$, $\alpha_0 = \alpha_N = (\pi/2N)$ are the quadrature coefficients.

Should we still consider the baby Dirichlet boundary-value problem for the operator L introduced in (5), the corresponding discrete problem would read:

$$\text{find } u_N \in \mathbb{P}_N, \quad u_N(-1) = u_-, \quad u_N(1) = u_+, \quad \text{s.t.} \\ (au_N', v_N')_N + (\beta u_N, v_N)_N + (\gamma u_N, v_N)_N = (f, v_N)_N, \\ \forall v_N \in \mathbb{P}_N^0 \quad (12)$$

where now $\mathbb{P}_N^0 = \{v_N \in \mathbb{P}_N \mid v_N(-1) = v_N(1) = 0\}$. This time, however, the expansion is made in terms of the nodal Lagrangian basis at Gauss-Lobatto nodes, that is, using instead of (9)

$$u_N(x) = \sum_{j=0}^N u_j \psi_j(x)$$

where ψ_j is the unique algebraic polynomial of degree N such that $\psi_j(x_i) = \delta_{ij}$, $\forall i, j = 0, \dots, N$.

One may show that

$$\psi_j(x) = \frac{-1}{N(N+1)} \cdot \frac{(1-x^2)}{(x-x_j)} \cdot \frac{L_N'(x)}{L_N'(x_j)}, \quad j = 0, \dots, N \quad (13)$$

The same approximation framework can be set up by replacing the Chebyshev polynomials with the Legendre polynomials $\{L_k, k = 0, 1, \dots\}$, which are orthogonal with respect to the traditional L^2 -scalar product (otherwise said with respect to the weight function $\omega = 1$).

The approximate problem still reads like (12); however, this time the nodes $\{x_j\}$ and the coefficients $\{\alpha_j\}$ are those of the (Legendre) Gauss-Lobatto integration.

The approach described above is named G-NI (Galerkin with Numerical Integration). A similar G-NI approach can be undertaken in several dimensions. For instance, consider

a second-order elliptic boundary-value problem

$$\begin{cases} Lu = f & \text{in } \Omega = (-1, 1)^d, \quad d = 2, 3 \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (14)$$

together with its weak form

$$\text{find } u \in V = H_0^1(\Omega) : a(u, v) = (f, v) \quad \forall v \in V \quad (15)$$

The bilinear form $a: V \times V \rightarrow \mathbb{R}$ is associated with the operator L ; for instance, if

$$Lu = -\text{div}(\alpha \nabla u) + \beta \cdot \nabla u + \gamma u, \quad \text{with } \alpha \geq \alpha_0 > 0 \quad (16)$$

then $a(u, v) = \int_{\Omega} (\alpha \nabla u \cdot \nabla v + \beta \cdot \nabla uv + \gamma uv)$.

Upon introducing the tensorized Legendre-Gauss-Lobatto quadrature nodes and coefficients $\mathbf{x}_k = (x_{k_1}, \dots, x_{k_d})$ and $a_k = a_{k_1}, \dots, a_{k_d}$ ($k_1 = 0, \dots, N$), the Legendre-Galerkin approximation of (15) with numerical integration (G-NI) becomes

$$\begin{aligned} \text{find } u_N \in V_N = \mathbb{P}_N^0 : \\ a_N(u_N, v_N) = (f, v_N)_N \quad \forall v_N \in V_N \end{aligned} \quad (17)$$

where \mathbb{P}_N is now the set of polynomials of degree $\leq N$ with respect to each of the independent variables, and \mathbb{P}_N^0 is its subspace made of those polynomials vanishing at $\partial\Omega$. Moreover,

$$(u, v)_N = \sum_k a_k u(\mathbf{x}_k) v(\mathbf{x}_k) \quad (18)$$

is the Gauss-Lobatto quadrature formula that approximates the scalar product (u, v) , while a_N is the discrete bilinear form that is obtained from a by replacing each scalar product (\cdot, \cdot) with $(\cdot, \cdot)_N$. Owing to the property that the quadrature formula (18) has degree of exactness $2N-1$, the Galerkin numerical integrated problem (17) can still be interpreted as a collocation method. Indeed, it follows from (17) that $L_N u_N = f$ at all internal nodes \mathbf{x}_k , where L_N is the approximation of L obtained by replacing each exact derivative by the derivative of the interpolant I_N at the Gauss-Lobatto nodes. The interpolation operator I_N is defined as follows: $I_N v(\mathbf{x}_k) = v(\mathbf{x}_k)$, $I_N v \in (\mathbb{P}_N)^d$, for all $v \in C^0(\bar{\Omega})$. Then, the operator approximating (16) is

$$L_N u_N = -\text{div}(I_N(\alpha \nabla u_N)) + \beta \cdot \nabla u_N + \gamma u_N$$

Existence and uniqueness of the solution of (18) follow from the assumption that $a_N(\cdot, \cdot)$ is a uniformly coercive form on the space $V \times V$, that is,

$$\begin{aligned} \exists \alpha^* > 0 \text{ independent of } N \text{ s.t.} \\ a_N(v_N, v_N) \geq \alpha^* \|v_N\|_{H^1(\Omega)}^2, \quad \forall v_N \in V_N \end{aligned} \quad (19)$$

This is the case for the problem at hand if, for example, β is constant and γ is nonnegative.

The convergence analysis of the G-NI approximation can be carried out by invoking the Strang Lemma for generalized Galerkin approximation. Precisely, the following error estimate holds:

$$\begin{aligned} \|u - u_N\| \leq \inf_{w_N \in V_N} \left[\left(1 + \frac{M}{\alpha^*} \right) \|u - w_N\| \right. \\ \left. + \frac{1}{\alpha^*} \sup_{v_N \in V_N \setminus \{0\}} \frac{a(w_N, v_N) - a_N(w_N, v_N)}{\|v_N\|} \right] \\ + \frac{1}{\alpha^*} \sup_{v_N \in V_N \setminus \{0\}} \frac{(f, v_N) - (f, v_N)_N}{\|v_N\|} \end{aligned}$$

where $\|\cdot\|$ is the norm of $H^1(\Omega)$ and M is the constant of continuity of the bilinear form $a(\cdot, \cdot)$.

Three sources contribute to the approximation error:

- the best approximation error, which can be immediately bounded by taking $w_N = I_{N-1}u$:

$$\inf_{w_N \in V_N} \|u - w_N\| \leq \|u - I_{N-1}u\|$$

- the error on the numerical quadrature, which can be bounded as follows:

$$\begin{aligned} \sup_{v_N \in V_N \setminus \{0\}} \frac{(f, v_N) - (f, v_N)_N}{\|v_N\|} \\ \leq C_2 (\|f - I_N f\|_{L^2(\Omega)} + \|f - P_{N-1} f\|_{L^2(\Omega)}) \end{aligned}$$

where $P_{N-1}f$ is the truncated Legendre series of f of order $N-1$;

- the error generated by the approximation of the bilinear form, on its hand, is less immediate to estimate. However, having chosen $w_N = I_{N-1}u$, which is a polynomial of degree $N-1$, using the degree of exactness of the quadrature formula and assuming that the coefficients of the operator are constant, one easily checks that $a(w_N, v_N) - a_N(w_N, v_N) = 0$, that is, this error is actually null. If the coefficients are nonconstant, one can control it in terms of the interpolation error measured in $H^1(\Omega)$.

We can conclude by taking advantage of the optimality of the truncation error in the L^2 -norm and that of the interpolation error in both the L^2 - and H^1 -norm:

$$\begin{aligned} \forall f \in H^r(\Omega), \quad r \geq 0, \quad \|f - P_N f\|_{L^2(\Omega)} \leq C_3 N^{-r} \|f\|_{H^r(\Omega)} \\ \forall g \in H^s(\Omega), \quad s \geq 1, \quad \|g - I_N g\|_{L^2(\Omega)} + \|g - I_N g\|_{H^1(\Omega)} \\ \leq C_4 N^{1-s} \|g\|_{H^s(\Omega)} \end{aligned}$$

Thus, we obtain that

$$\|u - u_N\| \leq C_5 (N^{-r} \|f\|_{H^r(\Omega)} + N^{1-s} \|u\|_{H^s(\Omega)})$$

provided u and f have the requested regularity.

A few comments on the implementation of the method are in order. The algebraic system associated with (17) reads $\mathbf{A}\mathbf{u} = \mathbf{f}$, where $a_{ij} = a_N(\psi_j, \psi_i)$, $f_i = (f, \psi_i)_N$, $\mathbf{u} = (u_i)$, $u_i = u_N(\mathbf{x}_i)$, and $\{\psi_j\}$ denote the Lagrangian basis functions of S_N^0 associated with all the nodal points $\{\mathbf{x}_i\}$. The matrix \mathbf{A} , which is nonsingular whenever (19) is fulfilled, is ill conditioned: indeed, there exist two constants C_1, C_2 such that

$$C_1 N^3 \leq \text{cond}(\mathbf{A}) \leq C_2 N^3$$

where $\text{cond}(\mathbf{A})$ is the (spectral) condition number of \mathbf{A} . The use of a preconditioned iterative procedure (e.g., the conjugate gradient when $\beta = 0$, or a Krylov iteration otherwise) is mandatory. A possible preconditioner is given by the diagonal of \mathbf{A} . This yields a preconditioned system whose condition number behaves like a constant times N^2 . A more drastic improvement would be achieved by taking as a preconditioner the matrix associated with the (piecewise-linear) finite element discretization of the operator (16) at the same Legendre-Gauss-Lobatto nodes. This is an optimal preconditioner as the condition number of the preconditioned system becomes independent of N .

Spectral methods based on algebraic polynomials have been discussed and analyzed in Canuto *et al.* (1988), Bernardi and Maday (1992), Bernardi and Maday (1997) and Guo (1998) (see Chapter 3, this Volume).

4 ALGEBRAIC EXPANSIONS ON TRIANGLES

Spectral methods for multidimensional problems rely their efficiency on the tensor product structure of the expansions they use. This feature naturally suggests the setting of the methods on patches of Cartesian products of intervals, such as squares or cubes, possibly after applying a smooth mapping. On the other hand, triangles, tetrahedra, prisms, and similar figures allow one to handle complex geometries in a more flexible way. So, a natural question arises: Can one match the advantages of a tensor product structure with those of a triangular geometry?

A positive answer to this question was given by Dubiner (1991), who introduced the concepts of collapsed Cartesian coordinate systems and warped tensor products. The method was further developed by Karniadakis and

Sherwin (1999). We now describe this approach in 2D, pointing to the latter reference for the 3D extensions. Let us introduce the reference triangle $\mathcal{T} = \{(x_1, x_2) \in \mathbb{R}^2 : -1 < x_1, x_2; x_1 + x_2 < 0\}$, as well as the reference square $\mathcal{Q} = \{(\xi_1, \xi_2) \in \mathbb{R}^2 : -1 < \xi_1, \xi_2 < 1\}$. The mapping

$$\begin{aligned} (x_1, x_2) \mapsto (\xi_1, \xi_2), \quad \xi_1 = 2 \frac{1+x_1}{1-x_2} - 1 \\ \xi_2 = x_2 \end{aligned} \quad (20)$$

is a bijection between \mathcal{T} and \mathcal{Q} . Its inverse is given by

$$\begin{aligned} (\xi_1, \xi_2) \mapsto (x_1, x_2), \quad x_1 = \frac{1}{2}(1+\xi_1)(1-\xi_2) \\ x_2 = \xi_2 \end{aligned}$$

Note that the mapping $(x_1, x_2) \mapsto (\xi_1, \xi_2)$ sends the ray in \mathcal{T} issuing from the upper vertex $(-1, 1)$ and passing through the point $(x_1, -1)$ into the vertical segment in \mathcal{Q} of equation $\xi_1 = x_1$. Consequently, the transformation becomes singular at the upper vertex, although it stays bounded therein. The Jacobian of the inverse transformation is given by $|\partial(x_1, x_2)/\partial(\xi_1, \xi_2)| = (1/2)(1-\xi_2)$. We term (ξ_1, ξ_2) the *collapsed Cartesian coordinates* of the point on the triangle whose regular Cartesian coordinates are (x_1, x_2) .

Denote by $\{P_k^{(a,b)}(\xi)\}$ the family of Jacobi polynomials of increasing degree $k \geq 0$, which form an orthogonal system with respect to the measure $(1-\xi)^a(1+\xi)^b d\xi$ in $(-1, 1)$ (note that $P_k^{(0,0)}(\xi)$ is the Legendre polynomial $L_k(\xi)$ introduced in the previous section). For $\mathbf{k} = (k_1, k_2)$, define the *warped tensor product* function on \mathcal{Q}

$$\Phi_{\mathbf{k}}(\xi_1, \xi_2) := \Psi_{k_1}(\xi_1) \Psi_{k_2}(\xi_2) \quad (21)$$

$$\text{where } \Psi_{k_1}(\xi_1) := P_{k_1}^{(0,0)}(\xi_1)$$

$$\Psi_{k_1, k_2}(\xi_2) := (1-\xi_2)^{k_1} P_{k_2}^{(2k_1+1,0)}(\xi_2) \quad (22)$$

which is a polynomial of degree k_1 in ξ_1 and $k_1 + k_2$ in ξ_2 . By applying the mapping (20), one obtains the function defined on \mathcal{T}

$$\begin{aligned} \varphi_{\mathbf{k}}(x_1, x_2) := \Phi_{\mathbf{k}}(\xi_1, \xi_2) = P_{k_1}^{(0,0)} \left(2 \frac{1+x_1}{1-x_2} - 1 \right) (1-x_2)^{k_1} \\ \times P_{k_2}^{(2k_1+1,0)}(x_2) \end{aligned} \quad (23)$$

It is easily seen that $\varphi_{\mathbf{k}}$ is a polynomial of global degree $k_1 + k_2$ in the variables x_1, x_2 . Furthermore, owing to the orthogonality of Jacobi polynomials, one has for

$\mathbf{k} \neq \mathbf{h}$

$$\begin{aligned} & \int_T \varphi_{\mathbf{k}}(x_1, x_2) \varphi_{\mathbf{h}}(x_1, x_2) dx_1 dx_2 \\ &= \frac{1}{2} \int_{-1}^1 P_{k_1}^{(0,0)}(\xi_1) P_{h_1}^{(0,0)}(\xi_1) d\xi_1 \\ & \times \int_{-1}^1 P_{k_2}^{(2k_1+1,0)}(\xi_2) P_{h_2}^{(2h_1+1,0)}(\xi_2) (1-\xi_2)^{k_1+h_1+1} d\xi_2 = 0 \end{aligned}$$

We conclude that the set $\{\varphi_{\mathbf{k}}: 0 \leq k_1, k_2; k_1 + k_2 \leq N\}$ is an orthogonal basis of modal type of the space $\mathcal{P}_N(T)$ of the polynomials of total degree $\leq N$ in the variables x_1, x_2 .

While orthogonality simplifies the structure of mass and stiffness matrices, it makes the enforcement of boundary conditions, or matching conditions between elements, uneasy. To overcome this difficulty, it is possible to modify the previous construction by building a new modal basis, say $\{\varphi_{\mathbf{k}}^*\}$, made of boundary functions (3 vertex functions plus $3(N-1)$ edge functions) and internal functions (bubbles). Each basis function retains the same 'warped tensor product' structure as above. Indeed, it is enough to replace in one dimension the Jacobi basis $P_k^{(\alpha,\beta)}(\xi)$ (with $\alpha = 0$ or $2k+1$) with the modified basis given by the two boundary functions $(1+\xi)/2$ and $(1-\xi)/2$ and the $N-1$ bubbles $(1+\xi)/2(1-\xi)/2 P_{k-1}^{(\alpha,\beta)}(\xi)$, $k = 1, \dots, N-1$. These univariate functions are then combined as in (21) to form the two-dimensional basis.

With such basis on hand, one can discretize a boundary-value problem by the Galerkin method with numerical integration (G-NI). To this end, one needs a high-precision quadrature formula on T . Since

$$\begin{aligned} & \int_T f(x_1, x_2) dx_1 dx_2 \\ &= \frac{1}{2} \int_{-1}^1 d\xi_1 \int_{-1}^1 F(\xi_1, \xi_2) (1-\xi_2) d\xi_2, \end{aligned}$$

it is natural to use a tensor product Gaussian formula in \mathcal{Q} for the measure $d\xi_1(1-\xi_2)d\xi_2$. This is obtained by tensorizing a $(N+1)$ -point Gauss-Lobatto formula for the measure $d\xi_1$ with a N -point Gauss-Radau formula for the measure $(1-\xi_2)d\xi_2$ with $\xi_2 = -1$ as integration knot (excluding the singular point $\xi_2 = 1$ from the integration knots makes the construction of the matrices easier). The resulting formula is exact for all polynomials in \mathcal{Q} of degree $\leq 2N-1$ in each variable ξ_1, ξ_2 ; hence, in particular, it is exact for all polynomials in T of total degree $\leq 2N-1$ in the variables x_1, x_2 . Note, however, that the number of quadrature nodes in T is $N(N+1)$, whereas the dimension of $\mathcal{P}_N(T)$ is $(1/2)(N+1)(N+2)$; thus, no basis in $\mathcal{P}_N(T)$ can be the Lagrange basis associated with the quadrature

nodes. This means that the G-NI method based on the quadrature formula described above cannot be equivalent to a collocation method at the quadrature points.

Finally, we observe that the G-NI mass and stiffness matrices on T can be efficiently built by exploiting the tensor product structure of both the basis functions and the quadrature points through the sum-factorization technique.

5 STOKES AND NAVIER-STOKES EQUATIONS

Spectral methods are very popular among the community of fluid-dynamicists. Owing to their excellent approximation properties, spectral methods can, in fact, provide very accurate simulations of complex flow patterns. However, special care is needed for the treatment of the incompressibility constraint. With the aim of simplification, let us first address the linear Stokes equations

$$\begin{cases} -\nu \Delta \mathbf{u} + \text{grad } p = \mathbf{f} & \text{in } \Omega \subset \mathbb{R}^d, d=2,3 \\ \text{div } \mathbf{u} = 0 & \text{in } \Omega \\ \mathbf{u} = 0 & \text{on } \partial\Omega \end{cases} \quad (24)$$

where $\nu > 0$ is the kinematic fluid viscosity, \mathbf{u} the fluid velocity, p the fluid pressure, and \mathbf{f} the volume forces.

A natural spectral G-NI discretization reads as follows: find $\mathbf{u}_N \in V_N$, $p_N \in Q_N$ such that

$$\begin{cases} ((\nu \nabla \mathbf{u}_N, \nabla \mathbf{v}_N))_N - (p_N, \text{div } \mathbf{v}_N)_N \\ = ((\mathbf{f}, \mathbf{v}_N))_N & \forall \mathbf{v}_N \in V_N \\ -(q_N, \text{div } \mathbf{v}_N)_N = 0 & \forall q_N \in Q_N \end{cases} \quad (25)$$

where $(\cdot, \cdot)_N$ is the discrete Gauss-Lobatto scalar product (18), while $((\cdot, \cdot))_N$ denotes its generalization to the case of vector functions. Moreover, $V_N = (\mathbb{P}_N^0)^d$ while Q_N is a polynomial space that needs to be chosen conveniently so as to satisfy the following Brezzi condition:

$$\begin{aligned} \exists \beta_N > 0: & \forall q_N \in Q_N, \exists \mathbf{v}_N \in V_N \text{ s.t. } (q_N, \text{div } \mathbf{v}_N)_N \\ & \geq \beta_N \|\mathbf{v}_N\|_{L^2(\Omega)} \|\mathbf{v}_N\|_{H^1(\Omega)} \end{aligned} \quad (26)$$

The violation of this condition, that is, the existence of nonconstant pressures $q_N \in Q_N$ such that $(q_N, \text{div } \mathbf{v}_N)_N = 0$, $\forall \mathbf{v}_N \in V_N$, implies the existence of *spurious pressure modes*, which pollute the computed pressure p_N .

The largest constant β_N , called the *inf-sup constant*, depends on the way Q_N is chosen, and has a special role in the analysis of the spectral approximation (25). Two choices are commonly proposed in practice. The first one is $Q_N = \mathbb{P}_{N-2} \cap L_0^2(\Omega)$, that is, the space of polynomials of degree $N-2$ with zero average. In that case, $\beta_N \simeq CN^{(1-d)/2}$.

An alternative approach consists of choosing $Q_N = \mathbb{P}_{[N]} \cap \mathbb{P}_{N-2} \cap L_0^2(\Omega)$, for some $\lambda: 0 < \lambda < 1$, where $[N]$ denotes the largest integer $\leq \lambda N$; in this case, $\beta_N \geq \beta > 0$.

The latter approach allows one to derive uniform stability and optimal error bounds for the approximate solution. In general, this occurs when β_N is uniformly bounded from below as N increases. In fact, using (26), one can obtain that

$$\nu \|\nabla \mathbf{u}_N\|_{(L^2(\Omega))^d} + \beta_N \|p_N\|_{L^2(\Omega)} \leq C \|\mathbf{f}\|_{(C^0(\bar{\Omega}))^d}$$

where C is a constant independent of N .

As a consequence, under the assumption (26), the error estimate on the velocity field is optimal, whereas the error on the pressure undergoes a loss of accuracy of order β_N^{-1} . For instance, in the case where $Q_N = \mathbb{P}_{N-2} \cap L_0^2(\Omega)$, the following error bound can be proven, provided the assumed regularity for the exact solution \mathbf{u} , p , and the forcing term \mathbf{f} holds for suitable values of $s \geq 1$ and $t \geq 0$:

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_N\|_{(H^1(\Omega))^d} + N^{(1-d)/2} \|p - p_N\|_{L^2(\Omega)} \\ \leq CN^{1-s} (\|\mathbf{u}\|_{(H^s(\Omega))^d} + \|p\|_{H^{t-1}(\Omega)}) + N^{-t} \|\mathbf{f}\|_{(H^t(\Omega))^d} \end{aligned}$$

Note that the $(N+1)^2$ Gauss-Lobatto nodes are used to interpolate the discrete velocity components, while the subset made of the $(N-1)^2$ interior Gauss-Lobatto nodes can be used to interpolate the discrete pressure. Alternatively, one could use a staggered grid made of the $(N-1)^2$ Gauss nodes for the pressure and change in (25) the discrete integrals $(p_N, \text{div } \mathbf{v}_N)_N$ and $(q_N, \text{div } \mathbf{u}_N)_N$ accordingly. This, however, would require interpolation between meshes, and, in this case, velocity and pressure feature nodal representations with respect to different sets of nodes.

The algebraic formulation of the discrete Stokes problem (25) yields the classical block structure matrix form

$$\begin{bmatrix} A & D^T \\ D & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ p \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ 0 \end{bmatrix} \quad (27)$$

where we have used test functions \mathbf{v}_N based on the Lagrangian polynomials of degree N of the velocity approximation, and test functions q_N based on the Lagrangian polynomials of degree $N-2$ (at the interior nodes) of the pressure approximation.

Upon eliminating (although only formally!) the \mathbf{u} vector, one obtains from (27) the reduced pressure system

$$S p = \mathbf{g}, \quad \text{with } S = D A^{-1} D^T \quad \text{and } \mathbf{g} = D A^{-1} \mathbf{f} \quad (28)$$

The pressure matrix S has $(N-1)^2$ rows and columns. It is symmetric; moreover, it is positive definite iff $\text{Ker } D^T = 0$, a condition that is equivalent to (26).

If we consider the generalized eigenvalue problem $S \mathbf{w} = \lambda M \mathbf{w}$, where M is the pressure mass matrix $(\psi_j, \psi_i)_N$, $((\psi_j))$ being the Lagrangian polynomials (of degree $\leq N-1$) associated with the interior Gauss-Lobatto nodes), then the maximum generalized eigenvalue λ_{\max} is uniformly bounded (from above) by the coercivity constant α of the discrete bilinear form $((\nabla \mathbf{u}_N, \nabla \mathbf{v}_N))_N$ (we can assume $\alpha = 1$ in the case on hand), whereas the minimum one λ_{\min} is proportional to β_N^2 . As a consequence, the condition number of the matrix $M^{-1}S$ is $\text{cond}(M^{-1}S) \sim \beta_N^{-2}$, thus $\sim N^{d-1}$ in the case of the $\mathbb{P}_N - \mathbb{P}_{N-2}$ discretization.

Since S is close to M (the discrete variational equivalent of the identity operator), M can serve as preconditioner for a conjugate gradient solution of (28). The corresponding PCG (Preconditioned Conjugate Gradient) method will converge in $O(N^{1/2})$ iterations for 2D problems and in $O(N)$ for 3D ones. In practice, however, the convergence is faster, as the previous estimate on the asymptotic behavior of β_N is too pessimistic.

Several kinds of generalizations are in order.

First of all, we mention that the Stokes system (24) could be reduced to a single (vector) equation by L^2 -projection upon the divergence-free subspace $V_{\text{div}} = \{\mathbf{v} \in (H_0^1(\Omega))^d | \text{div } \mathbf{v} = 0\}$:

$$\text{find } \mathbf{u} \in V_{\text{div}}: \quad \int_{\Omega} \nu \nabla \mathbf{u} \cdot \nabla \mathbf{v} = \int_{\Omega} \mathbf{f} \cdot \mathbf{v}, \quad \forall \mathbf{v} \in V_{\text{div}}$$

Since this is a well-posed elliptic problem, a unique velocity field can be obtained and, afterward, a unique pressure p can be recovered in $L_0^2(\Omega)$.

The simple structure of the reduced problem calls for a Galerkin (or G-NI) discretization. However, a computer implementation is far from being trivial, as one should construct a set of polynomial basis functions that are inherently divergence-free. This task has been successfully accomplished only for some specific boundary-value problems, for instance, when Ω is a cylindrical domain and Fourier expansion in the angular direction is combined with an expansion in terms of Chebyshev polynomials in both the longitudinal and the radial direction. A similar idea is behind the approach by Batcho and Karniadakis (1994) to generate eigenfunctions of a generalized Stokes operator and use them as polynomial divergence-free functions.

A different kind of generalization consists of using equal-order interpolation $\mathbb{P}_N - \mathbb{P}_N$ for both discrete velocity and pressure fields. However, this choice would give rise to a couple of subspaces, V_N and Q_N , which violate the Brezzi condition (26), yielding spurious pressure modes that swamp the physically relevant pressure. In line with what is nowadays common practice in the finite element community, Canuto and van Kemenade (1996) have proposed and analyzed a stabilization by bubble functions. The

idea consists in adding to $(\mathbb{P}_N^b)^d$ a supplementary space spanned by local polynomial functions having support in one small element called cell. In 2D, a cell is a quadrilateral whose four vertices are four neighboring Gauss-Lobatto points, whereas in 3D, it is a brick whose eight vertices are eight such points. The new velocity space is now given by $V_N = (\mathbb{P}_N^b)^d \oplus B_N^b$, where B_N^b denotes the space of bubble functions, while the pressure space is simply $Q_N = \mathbb{P}_N \cap L_0^2(\Omega)$.

After a careful analysis on the effect of the interaction of the local bubble functions with the global polynomials, and upon eliminating the bubble functions, contribution by static condensation, it is proven that the new stabilized discrete problem can be regarded as a Galerkin problem like (25); however, the continuity equation is modified by the presence of the additional term

$$\sum_C \tau_C (J_h x_N, J_h (\nabla(q_b)))_C$$

which plays the role of a stabilizing term to damp the oscillatory pressure modes. Here, C is a generic cell and $(\cdot, \cdot)_C$ is the $L^2(C)$ scalar product. Moreover, q_b is the (piecewise-linear) finite element interpolant of the test function q_N at the Gauss-Lobatto nodes, $r_N := -\nu \Delta u_N + \nabla p_N - f$ is the residual, J_h is the L^2 -projection operator into the space of piecewise constant functions on the cells. Finally, τ_C is the cell-stabilization parameter, which can be expressed in terms of the cell size h_C , the magnitude of the velocity field on C , and the fluid viscosity. Several expressions for τ_C are actually available based on alternative approaches that are residual-free.

The Navier-Stokes equations

$$\begin{cases} \partial_t u - \nu \Delta u + C(u) \\ + \text{grad } p = f & \text{in } \Omega \subset \mathbb{R}^d, d = 2, 3 \\ \text{div } u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (29)$$

differ from (24) because of the presence of the acceleration term $\partial_t u$ and the convective term $C(u)$. The latter can take the standard convective form $u \cdot \nabla u$; however, other expressions are used as well, such as the conservative form $\text{div}(uu)$ or the skew-symmetric form $(1/2)(u \cdot \nabla u + \text{div}(uu))$. The three forms are all equivalent for the continuous equations (with homogeneous Dirichlet boundary conditions) because of the incompressibility condition. However, this is no longer true at the discrete level. Indeed, the G-NI spectral discretization of the Navier-Stokes equations has different stability properties depending upon which form of $C(u)$ is employed.

For the time discretization of (29), fully implicit methods would produce a nonsymmetric, nonlinear system.

To avoid that, the convective term must be treated explicitly. One way is to combine backward-difference (BDF) discretization of linear terms with Adams-Bashforth (AB) discretization of the convective one. A classical recipe is the so-called BDF2/AB3, that is, the combination of the second-order BDF discretization with the third-order AB discretization:

$$\begin{aligned} & \left(\frac{3}{2\Delta t} M + A \right) u^{n+1} + D^T p^{n+1} \\ &= \frac{1}{\Delta t} M \left(2u^n - \frac{1}{2} u^{n-1} \right) + M f^{n+1} \\ & \quad - 2 \left(\frac{23}{12} C(u^n) - \frac{4}{3} C(u^{n-1}) + \frac{5}{12} C(u^{n-2}) \right) \end{aligned}$$

$$D u^{n+1} = 0$$

where M is now the velocity mass matrix, while A , D^T , and D are the matrices introduced before. To increase time accuracy, a BDF3 discretization is coupled with an extrapolation of the nonlinear term. This gives (Karniadakis, Israeli and Orszag, 1991)

$$\begin{aligned} & \left(\frac{11}{6\Delta t} M + A \right) u^{n+1} + D^T p^{n+1} \\ &= \frac{1}{\Delta t} M \left(3u^n - \frac{3}{2} u^{n-1} + \frac{1}{3} u^{n-2} \right) + M f^{n+1} \\ & \quad - \left(3C(u^n) - 3C(u^{n-1}) + C(u^{n-2}) \right) \end{aligned}$$

$$D u^{n+1} = 0$$

This scheme is third-order accurate with respect to Δt .

An extensive coverage of the spectral method for Navier-Stokes equations can be found in the books Canuto et al. (1988), Deville, Fischer and Mund (2002), and Peyret (2002). For their analysis, see also Bernardi and Maday (1992) and Bernardi and Maday (1997) (see Chapter 3, Volume 3, Chapter 9, this Volume).

6 ADVECTION EQUATIONS AND CONSERVATION LAWS

In order to illustrate spectral approximations to hyperbolic problems, we consider the linear and nonlinear 1D model equations $u_t + au_x = 0$ and $u_t + f(u)_x = 0$, supplemented by initial and appropriate boundary conditions. In addition to the standard issues related to spectral discretizations (efficient implementation, imposition of boundary conditions, stability, accuracy for smooth solutions), here we face a new problem. Indeed, the equation may propagate singularities along characteristics, or even (in the nonlinear case)

generate singularities from smooth initial data. So, the question arises: what is the interest of using high-order methods in such cases? We will answer this question in the second part of the present section.

For periodic problems, say in $(0, 2\pi)$, the Fourier-Galerkin method is the conceptually simplest choice: find $u_N = u_N(t) \in S_N$ such that

$$(u_{N,t} + au_{N,x}, \psi) = 0 \quad \text{or} \quad (u_{N,t} + f(u_N)_x, \psi) = 0, \quad \forall \psi \in S_N$$

Taking $\psi = u_N$, integrating by parts, and using periodicity, one obtains $(d/dt) \|u_N(t)\|_{L^2(0,2\pi)}^2 \leq K \|u_N(t)\|_{L^2(0,2\pi)}^2$ (with $K = \max_{0 \leq x \leq 2\pi} |a_x|$) for the linear advection equation, and $(d/dt) \|u_N(t)\|_{L^2(0,2\pi)}^2 = [F(u_N(t))]_0^{2\pi} = 0$ (where F denotes any primitive of f) for the conservation law. This proves the L^2 -stability of the approximation.

In terms of Fourier coefficients, the Galerkin method for the advection equation is equivalent to the set of ordinary differential equations

$$(u_{N,k})_t + (au_{N,x})_k = 0, \quad -\frac{N}{2} \leq k \leq \frac{N}{2} - 1$$

Setting for simplicity $b = u_{N,x}$, we have $(ab)_k = \sum_{h=-N/2}^{(N/2)-1} a_{k-h} b_h$. This is a family of convolution sums, which can be computed in $O(N^2)$ operations. A more efficient scheme consists of transforming back a and b in physical space, taking the pointwise product at the nodes $x_j = \pi j/N$, $j = 0, \dots, N-1$, and returning to Fourier space. Using the FFT, the full process costs $O(N \log N)$ operations. This is the *pseudospectral evaluation* of convolutions sums. There is an error involved, since one replaces the exact projection $P_N(ab)$ of ab upon S_N by its interpolant $I_N(ab)$ at the nodes. Such error, termed the *aliasing error*, is negligible if N is so large that the essential features of u are resolved. Otherwise, appropriate de-aliasing techniques can be applied, such as increasing the number of interpolation nodes.

This process applies to the conservation law as well, provided the nonlinearity is polynomial (as for Burgers's equation, $f(u_N) = (1/2)u_N^2$, or for the convective term $u_N \nabla u_N$ in the Navier-Stokes equations). It can be extended to the nonperiodic case by using the Chebyshev nodes $x_j = \cos \pi j/N$, $j = 0, \dots, N$.

The Fourier-Galerkin method with the pseudospectral evaluation of convolutions sums is nothing but the Galerkin method with numerical integration described in (6), or equivalently, the collocation method at the quadrature points

$$u_{N,j}^*(x_j) + a(x_j) u_{N,x}^*(x_j) = 0, \quad j = 0, \dots, N-1$$

Unless $a(x) \geq a > 0$ for all x , this scheme is (weakly) unstable due to the aliasing error. Writing the convective term in the skew-symmetric form

$$au_x = \frac{1}{2}(au)_x + \frac{1}{2}au_x - \frac{1}{2}a_x u \quad (30)$$

and applying pseudospectral derivatives, that is, the derivative of the interpolant, one recovers the same stability estimates as for the pure Galerkin method (in practice, such an expensive form is rarely necessary). Again, similar considerations apply in the nonlinear case as well.

We now turn to the discretization of nonperiodic problems, in the framework of Legendre methods. The advection equation is well-posed, provided we prescribe the solution, say $u(x_b) = g_b$, at the inflow points $x_b \in B_-$, where $B_\pm = \{x_b \in [-1, 1] : (\pm 1)a(x_b)n_b > 0\}$ with $n_b = x_b$. The most obvious way to account for the boundary conditions is to enforce them exactly (or strongly) in the discrete solution: $u_N \in \mathbb{P}_N$ satisfies $u_N(x_b) = g_b$, $\forall x_b \in B_-$. The corresponding Galerkin method is L^2 -stable. Indeed, assuming for simplicity $g_b = 0$, we take u_N itself as the test function and after integration by parts we get

$$\begin{aligned} & \frac{1}{2} \frac{d}{dt} \|u_N\|_{L^2(-1,1)}^2 - \frac{1}{2} K \|u_N\|_{L^2(-1,1)}^2 \\ & + \sum_{x_b \in B_+} a(x_b) n_b u_N^2(x_b) \leq 0 \end{aligned}$$

whence stability easily follows. A similar result holds for the Galerkin method with numerical integration (G-NI) at the Legendre-Gauss-Lobatto points, provided we use the skew-symmetric form (30) of the convective term. The G-NI scheme is equivalent to enforcing the equation at the internal nodes and at the noninflow boundary points ($x_b \notin B_-$).

A more flexible way to handle the boundary conditions, useful, for example, in domain decomposition and for systems of equations, is to enforce them in a weak sense. The rationale is that if stability holds then accuracy is assured, provided the boundary conditions are matched to within the same consistency error as for the equation in the interior. Thus, we seek $u_N = u_N(t) \in \mathbb{P}_N$ satisfying, for all $v_N \in \mathbb{P}_N$,

$$\begin{aligned} & (u_{N,t}, v_N)_N - (u_N, av_{N,x})_N + \sum_{x_b \in B_+} a(x_b) n_b u_N(x_b) v_N(x_b) \\ &= \sum_{x_b \in B_-} |a(x_b) n_b| g_b v_N(x_b) \end{aligned} \quad (31)$$

This G-NI formulation follows by integrating by parts the convective term (for simplicity, we assume a to be constant; otherwise, we use the skew-symmetric form (30)).

Choosing as v_N the polynomial Lagrange characteristic at each quadrature node, we see that the advection equation is enforced at all internal and noninflow nodes, whereas at the inflow nodes we have

$$u_{N,t}(x_b) + au_{N,x}(x_b) + \frac{1}{w_b} |a(x_b) n_b| (u_N(x_b) - g_b) = 0$$

Since $1/w_b \sim cN^2$ as $N \rightarrow +\infty$, this shows that the boundary condition is indeed enforced by a *penalty* method. The stability of the scheme (31) immediately follows by taking $v_N = u_N$. Stability is actually guaranteed even if we multiply each boundary term in (31) by any constant $\tau_b \geq 1/2$, thus enhancing the flexibility of the penalty method.

Spectral methods for linear advection equations are addressed by Gottlieb and Hesthaven (2001), Funaro (1997), and Fornberg (1996).

Let us now consider the nonlinear conservation law $u_t + f(u)_x = 0$. The stability (and convergence) of spectral discretizations is a much more delicate issue than that for the linear advection equation. Indeed, the equation may develop singular solutions at a finite time, which correspond to the accumulation of energy in the high-frequency modes or, equivalently, to the onset of oscillations around discontinuities (Gibbs phenomenon). The nonlinear mechanism may amplify the high-frequency components, leading to destructive instabilities (in stronger norms than L^2). On the other hand, oscillations should not be brutally suppressed: they are inherent to the high-order representation of discontinuous functions, and they may hide the correct information that allows the reconstruction of the exact solution. Thus, a good spectral discretization should guarantee enough stability while preserving enough accuracy. Furthermore, the discrete solution should converge to the physically relevant exact solution by fulfilling an appropriate entropy condition.

The mathematically most rigorous discretization that matches these requirements is the *spectral viscosity method* (see, e.g. Tadmor, 1998). In the Fourier–Galerkin context, it amounts to considering the modified equation

$$u_{N,t} + (P_N f(u_N))_x = \varepsilon_N (-1)^j D_x^2 (Q_m D_x^2 u_N)$$

where $\varepsilon_N \sim cN^{1-2s}$, $m = m_N \sim N^s$ for some $s < 1 - 1/(2s)$, and the Fourier coefficients of Q_m satisfy $Q_{m,k} = 0$ if $|k| \leq m$, $Q_{m,k} = 1 - (m/|k|)^{2s-1}$ if $|k| > m$. Thus, the s th order artificial viscosity is applied only to sufficiently high-frequency modes. For $s = 1$, one can prove that the solution is bounded in $L^\infty(0, 2\pi)$, it satisfies the estimate $\|u_N(t)\|_{L^2(0, 2\pi)} + \sqrt{\varepsilon_N} \|u_{N,x}(t)\|_{L^2(0, 2\pi)} \leq C \|u_N(0)\|_{L^2(0, 2\pi)}$, and it converges to the correct entropy solution.

A computationally simpler and widely used road to stabilization consists of *filtering* the spectral solution when advancing in time,

$$u_N(t) \mapsto \mathcal{F}_N u_N(t) = \sum_{k=-N/2}^{(N/2)-1} \sigma\left(\frac{2k}{N}\right) u_{N,k}(t) \exp(ikx)$$

where $\sigma = \sigma(\eta)$ is a smooth, even function satisfying $\sigma(0) = 1$, $\sigma^{(j)}(0) = 0$ for all j with $1 \leq j \leq s$, monotonically decreasing for $\eta > 0$ and vanishing (or being exponentially small) for $\eta > 1$. A popular choice is the exponential filter $\sigma(\eta) = \exp(-\alpha\eta^{2s})$. Interestingly, the effect of the spectral viscosity correction described above can be closely mimicked by applying the exponential filter with $\sigma(2k/N) = \exp(-\varepsilon_N Q_{m,k} k^2)$.

If the solution of the conservation law is piecewise analytic but discontinuous, its truncation $P_N u$ or its interpolation $I_N u$ are highly oscillatory around the singularities, and converge slowly ($O(N^{-1})$) to u away from them. However, they contain enough information to allow the reconstruction of the exact solution with exponential accuracy, away from the singularities, by a postprocessing as described below. It follows that the crucial feature of the discretization scheme is the capability of producing an approximation u_N , which is *spectrally close* to $P_N u$ or to $I_N u$. This is precisely what is obtained by the spectral viscosity method or by the equivalent filtering procedure.

Given $P_N u$ (similar considerations apply to $I_N u$), the postprocessing reconstruction may be *local* or *global*. In the former case, a spectrally accurate approximation of u at a point x_0 of analyticity is given by $u_N^*(x_0) = \int_{\beta} K_\alpha(x_0, y) \alpha(x_0 - y) P_N u(y) dy$, where $\beta = [\beta N]$ for some $\beta \in (0, 1)$, $K_\alpha(x, y)$ is, for each x , a v -degree polynomial approximation of the delta at x (e.g., for Fourier, $K_\alpha(x, y) = 1 + \sum_{n=2}^v \cos(x - y)$ is the Dirichlet kernel), whereas $q(\eta)$ is a C^∞ -localizer around $\eta = 0$. In the latter case, a spectrally accurate approximation of u on an interval $[a, b]$ of analyticity is given (Gottlieb and Shu, 1997) by the orthogonal projection of $P_N u$ upon $\mathbb{P}_s([a, b])$ (again $v = [\beta N]$) with respect to the weighted inner product $\int_a^b u(x) v(x) \omega_\alpha(x) dx$, with $\omega_\alpha(x) = ((x - a)(b - x))^{-1/2}$, which varies with N . The projection is computed via the Gegenbauer polynomials (i.e., the Jacobi polynomials $(P_k^{(\alpha-1/2, \alpha-1/2)})$ translated and scaled to $[a, b]$).

The reader can refer, for example, to Gottlieb and Shu (1997), Gottlieb and Tadmor (1984), and Tadmor (1998).

7 THE SPECTRAL ELEMENT METHOD

The spectral element method (SEM) represents another example of the Galerkin method. However, the finite

dimensional space is now made of piecewise algebraic polynomials of high degree on each element of a fixed partition of the computational domain. For a one-dimensional problem, such as, for example (2), we split $\Omega = (a, b)$ into a set of M disjoint intervals Ω_e , $e = 1, \dots, M$, whose end points are $a = \bar{x}_0 < \bar{x}_1 < \dots < \bar{x}_M = b$. Then we set

$$V_{N,M} = \{v \in C^0(\bar{\Omega}) : v|_{\Omega_e} \in \mathbb{P}_N, \quad \forall e = 1, \dots, M \\ v(a) = v(b) = 0\}$$

The approximation of (2) by SEM reads

$$\text{find } u_{N,M} \in V_{N,M} : a(u_{N,M}, v) = (f, v) \quad \forall v \in V_{N,M} \quad (32)$$

This approach shares the same structure as the p -version of the finite element method (FEM). As in the latter, the number M of subintervals is frozen, while the local polynomial degree (that is indicated by N in the SEM context and by p in the FEM context) is increased to improve accuracy. More precisely, if $h = (b - a)/M$ denotes the constant length of each subinterval, one has

$$\|u' - (\Pi_{N,M} u)'\|_{L^2(a,b)} + \frac{N}{h} \|u - \Pi_{N,M} u\|_{L^2(a,b)} \\ \leq C(s) h^{\min(N,s)} N^{-s} \|u'\|_{H^s(a,b)}, \quad s \geq 0 \quad (33)$$

where $\Pi_{N,M}$ is the SEM interpolant.

If u is arbitrarily smooth (s large), it is advantageous to keep h fixed and let $N \rightarrow \infty$.

Should the different degree of smoothness suggest the use of a nonuniform polynomial degree, another upper bound for the left-hand side of (33) is

$$\sum_{e=1}^M C_e h_e^{\min(N_e, s)} N_e^{-s} \|u'\|_{H^s(\Omega_e)}, \quad s \geq 1, \\ \forall e = 1, \dots, M$$

where N_e is the polynomial degree used in the e -th element Ω_e , and $H^{s+1/2}(\Omega_e)$ is the local smoothness of u in Ω_e .

SEM was first introduced by Patera (1984) for Chebyshev expansions, then generalized to the Legendre case by Y. Maday and A. Patera.

Both approaches (SEM and p -version of FEM) make use of a parental element, say $\hat{\Omega} = (-1, 1)$, on which the basis functions are constructed. However, the main difference lies in the way the basis functions are chosen (and therefore in the structure of the corresponding stiffness matrix).

FEMs of p -type are defined in terms of the Legendre polynomials $L_k(\xi)$ of degree k ($k = 2, \dots, p$), $\xi \in$

$\hat{\Omega}$. Precisely, the $p+1$ modal basis functions on $\hat{\Omega}$ are defined by

$$\varphi_1(\xi) = \frac{1-\xi}{2}, \quad \varphi_p(\xi) = \frac{1+\xi}{2}, \\ \varphi_k(\xi) = \sqrt{\frac{2k-1}{2}} \int_{-1}^{\xi} L_{k-1}(s) ds = \frac{1}{\sqrt{2(2k-1)}} \\ \times (L_k(\xi) - L_{k-2}(\xi)), \quad k = 2, \dots, p$$

The first two terms ensure C^0 continuity of the trial functions.

For the algebraic realization of SEM, nodal basis functions are those introduced in (13). Being associated with the special set of Legendre–Gauss–Lobatto nodes, once they are mapped on the current element $(\Omega_e, e = 1, \dots, M)$, they can be used to generate shape functions, then allow us to use LGL quadrature formulas for the evaluation of the entries of the stiffness and other matrices and the right-hand side. This is reflected by replacing (32) with the more interesting SEM-NI version:

$$\text{find } u_{N,M} \in V_{N,M} : \sum_{e=1}^M a_{N,\Omega_e}(u_{N,M}, v) = \sum_{e=1}^M (f, v)_{N,\Omega_e} \\ \forall v \in V_{N,M} \quad (34)$$

where $(u, v)_{N,M}$ is the Legendre–Gauss–Lobatto inner product (11) in Ω_e , $(u, v)_{N,\Omega_e} = \sum_{j=0}^N \alpha_j^e u(x_j^e) v(x_j^e)$, with $\alpha_j^e = \alpha_j [(b-a)/2]$, x_j^e is the correspondent of x_j in Ω_e . Moreover, $a_{N,\Omega_e}(u, v)$ is the elemental bilinear form.

Still considering the case of the differential operator (5) as an instance, we end up with the following form:

$$a_{N,\Omega_e}(u, v) = (au', v)_{N,\Omega_e} + (\beta u', v)_{N,\Omega_e} + (\gamma u, v)_{N,\Omega_e}$$

in analogy with the left-hand side of (12).

The multidimensional case can be addressed by first introducing the tensorized basis functions on the parental element $\hat{\Omega} = (-1, 1)^d$ ($d = 2, 3$), then mapping basis functions and nodal points on every element Ω_e (now a quadrilateral or parallelepipedal structure, possibly with curved edges or surfaces). The functional structure of our problem remains formally the same as in (34), and the kind of error estimate that can be achieved is similar. Obviously, this time $V_{N,M}$ is made of globally continuous functions that satisfy homogeneous Dirichlet boundary data (if any). They are obtained by joining the elemental functions that are the mapping of the nodal basis functions according to the transformation $T_e: \hat{\Omega} \rightarrow \Omega_e$ that maps the parental element $\hat{\Omega}$ into the current element Ω_e .

We refer to the seminal paper by Patera (1984) and to the books by Bernardi and Maday (1997), Deville,

Fisher and Mund (2002), Karniadakis and Sherwin (1999), and Schwab (1998).

8 THE MORTAR METHOD

This method has been introduced by Bernardi, Maday, and Patera (1994) with the aim of allowing spectral elements having different polynomial degrees or being geometrically nonconforming, and also to allow the coupling of the spectral (element) method with the finite element method. Its generality, however, goes beyond these two specific examples. Consider, for the sake of illustration, the Poisson problem with homogeneous Dirichlet conditions. The idea is to approximate its weak form (13) by the following discrete problem:

$$\text{find } u_\delta \in V_\delta : \sum_{i=1}^M \int_{\Omega_i} \nabla u_\delta \cdot \nabla v_\delta = \sum_{i=1}^M \int_{\Omega_i} f v_\delta \quad \forall v_\delta \in V_\delta \quad (35)$$

Here, $\delta > 0$ is a parameter describing the quality of the discretization, and V_δ is a finite dimensional space that approximates $H_0^1(\Omega)$ without being contained into $C^0(\bar{\Omega})$. More precisely, V_δ is a subspace of the following space:

$$Y_\delta := \{v_\delta \in L^2(\Omega) \mid v_{\delta|\Omega_i} \in Y_{i,\delta}, \quad i = 1, \dots, M\} \quad (36)$$

where, for each $i = 1, \dots, M$, $Y_{i,\delta}$ is a finite dimensional subspace of $H^1(\Omega_i)$; it can be either a finite element space, or a polynomial spectral (elements) space. In any case, no requirement of compatibility is made for the restriction of the functions of Y_δ on the element interface Γ .

Heuristically, the space V_δ will be made up of functions belonging to Y_δ that satisfy some kind of matching across Γ . Precisely, assuming for simplicity that there are only two elements, if $v_\delta \in V_\delta$ and $v_\delta^{(1)} \in Y_{1,\delta}$, $v_\delta^{(2)} \in Y_{2,\delta}$ denotes its restriction to Ω_1 and Ω_2 respectively for a certain fixed index i , the following integral matching conditions should be satisfied:

$$\int_\Gamma (v_\delta^{(1)} - v_\delta^{(2)}) \mu_\delta^{(i)} = 0 \quad \forall \mu_\delta^{(i)} \in \Lambda_\delta^{(i)} \quad (37)$$

where $\Lambda_\delta^{(i)}$ denotes the restriction to Γ of the functions of $Y_{i,\delta}$.

If we take $i = 2$ in (37), this amounts to letting Ω_1 play the role of master and Ω_2 that of slave, and (37) has to be intended as the way of generating the value of $v_\delta^{(2)}$ once $v_\delta^{(1)}$ is available. The alternative way, that is, taking $i = 1$ in (37) is also admissible. Depending upon the choice of index i made in (37), the method will produce different solutions.

The mathematical rationale behind the choice of the matching condition (37) (rather than a more 'natural' condition of pointwise continuity at one set of grid nodes on Γ) becomes clear from the convergence analysis for problem (35).

With this aim, we introduce

$$\|v\|_s := (\|v\|_{0,\Omega}^2 + \|\nabla v\|_{0,\Omega_1}^2 + \|\nabla v\|_{0,\Omega_2}^2)^{1/2} \quad (38)$$

which is a norm (the 'graph' norm) for the Hilbert space

$$H_s := \{v \in L^2(\Omega) \mid v_{|\Omega_1} \in H^1(\Omega_1), v_{|\Omega_2} \in H^1(\Omega_2)\} \quad (39)$$

Owing to the Poincaré inequality, we have that

$$\sum_{i=1}^2 \int_{\Omega_i} |\nabla v_\delta|^2 \geq \alpha_\delta \|v_\delta\|_s^2 \quad \forall v_\delta \in V_\delta \quad (40)$$

whence the discrete problem (35) admits a unique solution by a straightforward application of the Lax–Milgram lemma.

For any $v_\delta \in V_\delta$, we now have

$$\begin{aligned} \alpha_\delta \|u_\delta - v_\delta\|_s^2 &\leq \sum_{i=1}^2 \int_{\Omega_i} |\nabla(u_\delta - v_\delta)|^2 \\ &= \sum_{i=1}^2 \int_{\Omega_i} \nabla u_\delta \cdot \nabla(u_\delta - v_\delta) - \sum_{i=1}^2 \int_{\Omega_i} \nabla v_\delta \cdot \nabla(u_\delta - v_\delta) \\ &= \sum_{i=1}^2 \int_{\Omega_i} f(u_\delta - v_\delta) - \sum_{i=1}^2 \int_{\Omega_i} \nabla v_\delta \cdot \nabla(u_\delta - v_\delta) \end{aligned} \quad (41)$$

Replacing f by $-\Delta u$ and integrating by parts on each Ω_i , we obtain

$$\begin{aligned} \sum_{i=1}^2 \int_{\Omega_i} f(u_\delta - v_\delta) &= \sum_{i=1}^2 \int_{\Omega_i} \nabla u \cdot \nabla(u_\delta - v_\delta) \\ &\quad - \int_\Gamma \frac{\partial u}{\partial n} [(u_\delta - v_\delta)^{(1)} - (u_\delta - v_\delta)^{(2)}] \end{aligned} \quad (42)$$

(here, $(\partial/\partial n)$ is the normal derivative on Γ pointing into Ω_2).

Denoting by

$$[v_\delta]_\Gamma := v_\delta^{(1)}|_\Gamma - v_\delta^{(2)}|_\Gamma$$

the jump across Γ of a function $v_\delta \in V_\delta$, from (41) and (42), we have that

$$\alpha_\delta \|u_\delta - v_\delta\|_s^2 \leq \|u - v_\delta\|_s \|u_\delta - v_\delta\|_s + \left| \int_\Gamma \frac{\partial u}{\partial n} [u_\delta - v_\delta]_\Gamma \right|$$

and also

$$\|u_\delta - v_\delta\|_s \leq \frac{1}{\alpha_\delta} \left(\|u - v_\delta\|_s + \sup_{w_\delta \in V_\delta} \left| \int_\Gamma \frac{\partial u}{\partial n} [w_\delta]_\Gamma \right| \right)$$

By the triangle inequality

$$\|u - u_\delta\|_s \leq \|u - v_\delta\|_s + \|u_\delta - v_\delta\|_s$$

we then obtain the following inequality for the error $u - u_\delta$:

$$\begin{aligned} \|u - u_\delta\|_s &\leq \left(1 + \frac{1}{\alpha_\delta} \right) \inf_{v_\delta \in V_\delta} \|u - v_\delta\|_s \\ &\quad + \frac{1}{\alpha_\delta} \sup_{w_\delta \in V_\delta} \left| \int_\Gamma \frac{\partial u}{\partial n} [w_\delta]_\Gamma \right| \end{aligned} \quad (43)$$

The approximation error of (35) is therefore bounded (up to a multiplicative constant) by the best approximation error (i.e., the distance between the exact solution u and the finite dimensional space V_δ) plus an extra error involving interface jumps. The latter would not appear in the framework of classical Galerkin approximation (like the SEM), and is the price to pay for the violation of the conforming property; that is, for the fact that $V_\delta \not\subset H_0^1(\Omega)$.

The error estimate (43) is optimal if each one of the two terms on the right can be bounded by the norm of local errors arising from the approximations in Ω_1 and Ω_2 , without the presence of terms that combine them in a multiplicative fashion. In this way, we can take advantage of the local regularity of the exact solution as well as the approximation properties enjoyed by the local subspaces $Y_{i,\delta}$ of $H^1(\Omega_i)$.

To generate a nodal basis for the finite dimensional space V_δ , we can proceed as follows. For $i = 1, 2$, let us denote by N_i the set of nodes in the interior of Ω_i , and by $N_\Gamma^{(i)}$ the set of nodes on Γ , whose cardinality will be indicated by N_i and $N_\Gamma^{(i)}$, respectively. Note that, in general, $N_\Gamma^{(1)}$ and $N_\Gamma^{(2)}$ can be totally unrelated.

Now, denote by $\{\varphi_k^{(i)}\}$, $k = 1, \dots, N_i$, the Lagrange functions associated with the nodes of N_i ; since they vanish on Γ , they can be extended by 0 in Ω_2 . These extended functions are denoted by $\{\tilde{\varphi}_k^{(i)}\}$, and can be taken as a first set of basis functions for V_δ .

Symmetrically, we can generate as many basis functions for V_δ as the number of nodes of N_2 by extending by 0 in Ω_1 the Lagrange functions associated with these nodes. These new functions are denoted by $\{\tilde{\varphi}_k^{(2)}\}$, $k' = 1, \dots, N_2$.

Finally, always supposing that Ω_1 is the master domain and Ω_2 its slave, for every Lagrange function $\{\varphi_m^{(1)}\}$ in

$\bar{\Omega}_1$, $m = 1, \dots, N_\Gamma^{(1)}$, we obtain a basis function $\{\tilde{\varphi}_{m,\Gamma}\}$ as follows:

$$\tilde{\varphi}_{m,\Gamma} := \begin{cases} \varphi_m^{(1)} & \text{in } \bar{\Omega}_1 \\ \tilde{\varphi}_{m,\Gamma}^{(2)} & \text{in } \bar{\Omega}_2 \end{cases}$$

where

$$\tilde{\varphi}_{m,\Gamma}^{(2)} := \sum_{j=1}^{N_\Gamma^{(2)}} \xi_j \varphi_{j,\Gamma}^{(2)}$$

$\varphi_{j,\Gamma}^{(2)}$ are the Lagrange functions in $\bar{\Omega}_2$ associated with the nodes of $N_\Gamma^{(2)}$, and ξ_j are unknown coefficients that should be determined through the fulfillment of the matching equations (37). Precisely, they must satisfy

$$\int_\Gamma \left(\sum_{j=1}^{N_\Gamma^{(2)}} \xi_j \varphi_{j,\Gamma}^{(2)} - \varphi_m^{(1)} \right) \varphi_k^{(2)} = 0 \quad \forall k = 1, \dots, N_\Gamma^{(2)} \quad (44)$$

A basis for V_δ is therefore provided by the set of all functions $\{\tilde{\varphi}_k^{(1)}\}$, $k' = 1, \dots, N_1$, $\{\tilde{\varphi}_k^{(2)}\}$, $k'' = 1, \dots, N_2$, and $\{\tilde{\varphi}_{m,\Gamma}\}$, $m = 1, \dots, N_\Gamma^{(1)}$.

Remark. In the mortar method, the interface matching is achieved through a L^2 -interface projection, or, equivalently, by equating first-order moments, thus involving computation of interface integrals. In particular, from equations (37), we have to evaluate two different kinds of integrals (take, for instance, $i = 2$):

$$I_{12} := \int_\Gamma v_\delta^{(1)} \mu_\delta^{(2)}, \quad I_{22} := \int_\Gamma v_\delta^{(2)} \mu_\delta^{(2)}$$

The computation of I_{22} raises no special difficulties, because both functions $v_\delta^{(2)}$ and $\mu_\delta^{(2)}$ live on the same mesh, the one inherited from Ω_2 . On the contrary, $v_\delta^{(1)}$ and $\mu_\delta^{(2)}$ are functions defined on different domains, and the computation of integrals like I_{12} requires proper quadrature rules. This process needs to be done with special care, especially for three-dimensional problems, for which subdomain interfaces are made up of faces, edges, and vertices; otherwise the overall accuracy of the mortar approximation could be compromised.

The discrete problem (35) can also be reformulated as a saddle point problem of the following form:

find that $u_\delta \in Y_\delta$, $\lambda_\delta \in \Lambda_\delta^{(2)}$ such that

$$\begin{cases} a(u_\delta, v_\delta) + b(v_\delta, \lambda_\delta) = \sum_{i=1}^2 \int_{\Omega_i} f v_\delta^{(i)} \quad \forall v_\delta \in Y_\delta \\ b(u_\delta, \mu_\delta) = 0 \quad \forall \mu_\delta \in \Lambda_\delta^{(2)} \end{cases}$$

where

$$a(w_h, v_h) := \sum_{i=1}^2 \int_{\Omega_i} \nabla w_h^{(i)} \cdot \nabla v_h^{(i)}$$

$$b(v_h, \mu_h) := \int_{\Gamma} (v_h^{(1)} - v_h^{(2)}) \mu_h$$

In this system, λ_h plays the role of the Lagrange multiplier associated with the 'constraint' (37).

Denoting by φ_j , $j = 1, \dots, N_1 + N_2 + N_{\Gamma}^{(1)} + N_{\Gamma}^{(2)}$, a basis of Y_h and by ψ_l , $l = 1, \dots, N_{\Gamma}^{(2)}$, a basis of $\Lambda_h^{(2)}$, we introduce the matrices

$$A_{ij} := a(\varphi_j, \varphi_i), \quad B_{li} := b(\varphi_i, \psi_l)$$

Defining by \mathbf{u} and $\boldsymbol{\lambda}$ the vectors of the nodal values of u_h and λ_h , respectively, and by \mathbf{f} the vector whose components are given by $\sum_{i=1}^2 (f, \varphi_i^{(j)})_{\Omega_i}$, $s = 1, \dots, N_1 + N_2 + N_{\Gamma}^{(1)} + N_{\Gamma}^{(2)}$, we have the linear system

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix}$$

The matrix A is block-diagonal (with one block per subdomain Ω_i), each block corresponding to a problem for the Laplace operator with a Dirichlet boundary condition on $\partial\Omega_i \cap \partial\Omega$ and a Neumann boundary condition on $\partial\Omega_i \setminus \partial\Omega$.

After elimination of the degrees of freedom internal to the subdomains, the method leads to the reduced linear system (still of a saddle point type)

$$\begin{bmatrix} S & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_{\Gamma} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_{\Gamma} \\ \mathbf{0} \end{bmatrix}$$

where the matrix S is block-diagonal, C is a jump operator, \mathbf{u}_{Γ} is the set of all nodal values at subdomain interfaces, and \mathbf{g}_{Γ} is a suitable right-hand side.

This system can be regarded as an extension of the Schur complement system to nonconforming approximation (the Lagrange multiplier λ indeed accounts for nonmatching discretization at subdomain interfaces). In fact, the i th block of S is the analogue of $\Sigma_{i,k}$, and corresponds to a discretized Steklov–Poincaré operator on the subdomain Ω_i .

Remark. All results cited in this note can be recovered from the books and general articles that are quoted in the References.

REFERENCES

- Batcho PF and Karniadakis GE. Generalized stokes Eigen functions: a new trial basis for the solution of incompressible Navier–Stokes equations. *J. Comput. Phys.* 1994; 115:121–146.
- Bernardi C and Maday Y. *Approximations Spectrales de Problèmes aux Limites Elliptiques*. Springer-Verlag: Paris, 1992.
- Bernardi C and Maday Y. Spectral methods. In *Handbook of Numerical Analysis, Volume V of Techniques of Scientific Computing*, Ciarlet PJ, Lions JL (eds). North Holland: Amsterdam, 1997; 209–486.
- Bernardi C, Maday Y and Patera AT. A new nonconforming approach to domain decomposition: the mortar element method. In *Nonlinear Partial Differential Equations and their Applications*, Collège de France Seminar, Vol. XI, Brezis H, Lions JL (eds). Longman Group: Harlow, 1994; 13–51.
- Canuto C and van Kemenaede V. Bubble stabilized spectral methods for the incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech. Eng.* 1996; 135:35–61.
- Canuto C, Hussaini MY, Quarteroni A and Zang TA. *Spectral Methods in Fluid Dynamics*. Springer-Verlag: Berlin, Heidelberg, 1988.
- Deville MO, Fischer PF and Mund EH. *High-Order Methods for Incompressible Fluid Flow*. Cambridge University Press: Cambridge, 2002.
- Dubiner M. Spectral methods on triangles and other domains. *J. Sci. Comput.* 1991; 6:345–390.
- Fornberg BA. *Practical Guide to Pseudospectral Methods*. Cambridge University Press: Cambridge, 1996.
- Funaro D. *Spectral Elements for Transport-Dominated Equations*. Springer-Verlag: Berlin, Heidelberg, 1997.
- Gottlieb D and Hesthaven JS. Spectral methods for hyperbolic problems. *J. Comput. Appl. Math.* 2001; 128:83–131.
- Gottlieb D and Orszag SA. *Numerical Analysis for Spectral Methods: Theory and Applications*. SIAM-CBMS: Philadelphia, 1977.
- Gottlieb D and Shu C-W. On the Gibbs phenomenon and its resolution. *SIAM Rev.* 1997; 39:644–668.
- Gottlieb D and Tadmor E. Recovering pointwise values of discontinuous data with spectral accuracy. *Progress and Supercomputing in CFD*. Birkhäuser: Boston, 1984; 357–375.
- Guo B-Y. *Spectral Methods and their Applications*. World Scientific Publishing: Singapore, 1998.
- Karniadakis GE and Sherwin SJ. *Spectral hp Element Method for CFD*. Oxford University Press: New York, 1999.
- Karniadakis GE, Israeli M and Orszag SA. High-order splitting methods for incompressible Navier–Stokes equations. *J. Comput. Phys.* 1991; 97:414–443.
- Patera AT. A spectral element method for fluid dynamics: laminar flow in a channel expansion. *J. Comput. Phys.* 1984; 54:468–488.
- Peyret R. *Spectral Methods for Incompressible Viscous Flow*. Springer-Verlag: New York, 2002.
- Schwab Ch. *p- and hp- Finite Element Methods*. Oxford Science Publications: New York, 1998.
- Tadmor E. Approximate solutions of nonlinear conservation laws. In *Advanced Numerical Approximations of Nonlinear Hyperbolic Equations*, Quarteroni A (eds). Springer: Berlin, 1998; 1–130.

Chapter 7

Adaptive Wavelet Techniques in Numerical Simulation

Albert Cohen¹, Wolfgang Dahmen² and Ronald DeVore³

¹ Université Pierre et Marie Curie, Paris, France

² RWTH Aachen, Germany

³ University of South Carolina, Columbia, SC, USA

| | |
|---|-----|
| 1 Introduction | 157 |
| 2 Wavelets | 160 |
| 3 Evolution Problems – Compression of Flow Fields | 169 |
| 4 Boundary Integral Equations – Matrix Compression | 175 |
| 5 A New Adaptive Paradigm | 181 |
| 6 Construction of Residual Approximations and Complexity Analysis | 187 |
| Acknowledgment | 195 |
| Notes | 195 |
| References | 195 |

1 INTRODUCTION

Increasingly realistic models in computational mechanics and the search for more and more accurate simulations place continuously growing demands on computation that surpass the ongoing increase of computing power. Thus, paradoxically, these finer models might be of limited use in the absence of new computational strategies. One promising, emerging strategy is to dynamically adapt discretizations in the course of the computational solution process. Adaptive

strategies of this type have been observed to reduce the complexity of computational problems arising in large scale numerical simulation. Therefore, adaptivity provides an enormous potential for advancing the frontiers of computability. By bringing more and more complex tasks into reach, it offers, in the long run, better and better access to physical phenomena through a powerful numerical microscope. On the other hand, to advance these techniques to their natural fruition requires an understanding of the power of adaptivity vis-a-vis traditional methods of computation. This includes clarifying the optimal performance that can be expected from adaptive methods and how this compares with the performance using nonadaptive techniques.

This chapter describes adaptive numerical strategies in the context of multiscale decompositions using wavelet bases. In addition to formulating adaptive strategies to be used in a variety of settings, the chapter will provide an a priori analysis of the computational efficiency of adaptive methods. This will delineate the advantages of adaptive strategies versus standard computational methods.

Adaptivity takes a variety of forms distinguished by their principal goals. In many applications, one is not interested in the complete solution of a given problem but only in certain local functionals of an object that may be globally defined like the solution of a boundary value problem. In this case, an adaptive discretization reflects how much has to be paid to the global character of the object when trying to recover local information about it. However, this is

not the direction of the present chapter. Instead, this chapter focuses on recovering the *whole* object in question. In the context of fluid mechanics, this may mean recovering the vortices in the wake of an airfoil, or the interaction of shocks even at some distance of the airfoil, or the recovery of a full stress field, or eventually, understanding more about developing turbulence. The objective is to develop numerical techniques that are able to extract information within desired error tolerances at minimal cost. This means that searched-for-quantities like pressure or velocity are to be recovered within some accuracy tolerance, for example, with respect to some norm. This should be done at the expense of a number of degrees of freedom that remains proportional to the minimal number of degrees of freedom (in a certain discretization framework) needed to approximate the object based on full information within the desired target accuracy. From the mathematical point of view, it is not clear beforehand at all whether this is possible solely based on a posteriori information acquired during a solution process. We shall indicate in this chapter an affirmative answer for a wide range of problems arising in engineering applications (see Cohen, Dahmen and DeVore, 2001, 2002a,b,c).

Our approach involves expansions of functions into *wavelet bases*. In such expansions, the wavelet coefficients encode *detail information* that has to be added when progressing to higher levels of resolution of the underlying function. These coefficients convey local structural information such as the regularity of the expanded function. The decomposition naturally breaks the function into different characteristic length scales. A central question in many dynamical simulation tasks concerns the interaction of these different scales. As we shall show, wavelet analysis offers a promising way to describe the behavior of contributions from different length scales under *nonlinear mappings*. We shall see that wavelet expansions offer quantitative ways of estimating nonlinear effects that appear, for example, in the Navier Stokes equations. Moreover, we shall point out how such an analysis aids the adaptive solution process. This already indicates the marriage between the analysis and numerical resolution of a problem facilitated by wavelet concepts. Therefore, the understanding of these concepts and their potential requires a certain amount of functional analysis as will be described in this chapter.

We do not attempt to give an exhaustive overview of wavelet analysis pertinent to computational mechanics issues. Nor will the topics presented here be treated in a self-contained way. Both would be far beyond the scope of this chapter. Rather, we shall focus on presenting some concepts and ideas, which in our opinion best reflect the potential of wavelets, thereby offering some orientation that could be complemented by the extensive list of references.

The following surveys and text books are recommended as sources of more detailed expositions: Cohen, 2000, 2003; Dahmen, 1997, 2001; DeVore, 1998.

The organization of the material is in some sense 'two-dimensional'. Most simulation tasks are based on continuous mathematical models formulated in terms of integral or (partial) differential equations. The 'first dimension' is to group the different concepts with respect to the following two major problem classes. The first one concerns *evolution problems*

$$\partial_t u = \mathcal{E}(u) \quad (1)$$

together with initial and boundary conditions. The second class concerns stationary problems

$$\mathcal{R}(u) = 0 \quad (2)$$

which are usually given in variational form. The scope of problems covered by (2) will be illustrated by a list of examples including mixed formulations and nonlinear problems. Of course, there is no clear dividing line. For instance, an implicit time discretization of a parabolic evolution problem leads to a family of problems of the type (2). The problems grouped under (2) are typically elliptic (in the sense of Agmon–Douglis–Nirenberg) for which *Hilbert space* methods are appropriate. In contrast, we focus under (1) on nonlinear *hyperbolic* problems. It will be seen that the respective concepts are quite different in nature. The nature of the relevant function spaces, for example, L_1 which admits no unconditional basis, causes an impediment to exploiting the full potential of wavelets.

The 'second dimension' of organization concerns the way wavelet features are exploited. In Section 2, we review briefly the main features that drive wavelets as analysis and discretization tools. Aside from transform mechanisms, these are the *locality* (in physical and frequency domain), the *cancellation properties*, and the *norm equivalences* between function and sequence spaces. The latter facilitates a stable coupling of the continuous and the discrete world. Together with the first two features, this is also fundamental for fast numerical processing.

In Section 3, these features are applied to (1). The primary focus of adaptivity here is the *sparse approximation* of the unknown solution, mainly owing to the cancellation properties. In this context, wavelets are *not* used as stand-alone tools but are rather combined with conventional finite volume discretizations. The numerical approximation represented by arrays of cell averages is compressed in a manner similar to image compression. This amounts to a *perturbation* analysis in which one seeks a significant data compression while preserving essentially the accuracy

of the underlying reference discretization for a *fixed* level of resolution. The approach and the performance of such schemes are illustrated by some numerical examples concerning aerodynamical applications.

The remainder of the chapter is concerned with the problem class (2). Section 4 deals with *global operators* represented here by the classical boundary integral equations. Now the above-mentioned main features of wavelets are used mainly to obtain *sparse approximations of operators*. This time the elliptic nature of the problem allows one to formulate stable Galerkin discretizations. When using finite elements or boundary elements, the resulting stiffness matrices are densely populated and, depending on the operator, are increasingly ill-conditioned when the mesh size decreases. In wavelet formulations, the norm equivalences and cancellation properties are used to show that the stiffness matrices can be replaced by sparse well-conditioned ones without sacrificing discretization error accuracy. This allows one to solve such problems in linear time. Again, this is essentially a perturbation approach in which this time sparse approximations apply to the operator and not to the function. Adaptivity refers here primarily to the quadrature used to compute the compressed stiffness matrices with a computational effort that stays proportional to the problem size. As for the current state of the art, we refer to Dahmen, Harbrecht and Schneider (2002), Harbrecht (2001), and the references cited there.

In Section 5, we introduce a *new algorithmic paradigm* that emerges from exploiting both, the sparse approximation of functions and the sparse representation of operators together. It aims at intertwining in some sense the analysis and resolution aspects of wavelet concepts as much as possible. Here are the main conceptual pillars of this approach:

A transform point of view: Many studies of wavelet methods for the numerical solution of PDEs are very similar in spirit to classical finite element discretizations where the trial spaces are spanned by *finite* collections of wavelets. This has so far dominated the use of wavelets in the context of boundary integral equations and is the point of view taken in Section 4. However, this does not yet fully exploit the potential of wavelets. In fact, similar to classical Fourier methods, wavelets can be used to formulate *transform methods* that are best explained in the context of variational formulations of (linear or nonlinear) operator equations like boundary value problems or boundary integral equations. Unlike finite volume or finite element schemes, wavelet bases can be used to transform the original variational problem into an *equivalent* problem over ℓ_2 , the space of square summable sequences indexed by the wavelet basis. Moreover, when the wavelets are correctly chosen in

accordance with the underlying problem, the transformed (still infinite dimensional) problem is now well-posed in a sense to be made precise later. We shall point out now the main principles along these lines.

Staying with the infinite dimensional problem: In many cases, the underlying infinite dimensional problem, for example, a PDE, is fairly well understood. In mathematical terms, this means that when formulated as an operator equation, the operator is known to be boundedly invertible as a mapping from a certain function space into its dual, which is another way of saying that the problem is well posed in a certain topology – there exists a unique solution, which depends continuously on the data in the topology given by that function space.

When transforming to the wavelet domain, the properties of the operator are inherited by the transformed operator which now acts on sequence spaces. The main point we wish to stress is that the original infinite dimensional problem is often better understood than specific discretized finite dimensional versions and therefore there is an advantage in delaying the movement to finite discretizations as long as possible. A classical example of this is the Stokes problem in which a positive definite quadratic functional is minimized under the divergence constraint and thus has *saddle point* character. The Stokes problem is well posed in the above sense for the right pairs of function spaces for the velocity and pressure component (see e.g. Brezzi and Fortin, 1991; Girault and Raviart, 1986). It is well-known that Galerkin discretizations, however, may very well become unstable unless the trial spaces for velocity and pressure satisfy a compatibility condition called the Ladyženskaya–Babuška–Brezzi (LBB) condition. For the Stokes problem, this is well understood but in other situations, as in many physically very appropriate *mixed formulations*, finding stable pairs of trial spaces is a more delicate task. So, in some sense one may run into self-inflicted difficulties when turning to finite discretizations even though the original problem is well behaved. Is there an alternative?

Stabilizing effects of adaptivity: The very fact that, unlike conventional schemes, a suitable wavelet basis captures the complete infinite dimensional problem and puts it into a well-conditioned format over ℓ_2 can be used to avoid fixing any finite dimensional discretization. Instead the well-posedness offers ways of formulating an iterative scheme for the *full infinite dimensional problem* that converges (conceptually) with a fixed error reduction per step. Only after this infinite dimensional analysis is complete do we enter the numerical stage by applying the involved infinite dimensional (linear and also nonlinear) operators *adaptively* within suitable stage-dependent dynamically updated

error tolerances. Roughly speaking, this numerical approach inherits the well-posedness of the original problem and allows us to avoid imposing compatibility conditions such as LBB.

Adaptive evaluation of operators: A central issue is then to actually realize concrete adaptive evaluation schemes for relevant operators and to analyze their computational complexity. We shall engage this issue for both linear and nonlinear examples. While at the first glance nonlinear operators seem to interfere in an essential way with wavelet concepts (as they do with regard to the Fourier techniques), we claim that they offer particularly promising perspectives in this regard. In conventional discretizations, the image of a current adaptive approximation under a nonlinear mapping is usually discretized on the same mesh. However, a singularity, which may cause an adaptive local refinement, is often severely affected by a nonlinearity so that this mesh might no longer be optimal. In the present framework, the adaptive evaluation will be seen to generate at each stage the right choice of degrees of freedom for the image of the nonlinearity; see Section 6.2. This is based on quantitative estimates on the interaction of different length scales under nonlinear mappings.

It would be far beyond the scope of this chapter to address any of the above issues in complete detail. Instead, our presentation will be more of an overview of this subject which should serve to orient the reader to the essential concepts and point of views. The interested reader will then find extensive references for further reading.

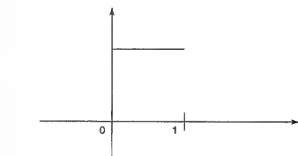
2 WAVELETS

In this section, we give a brief overview of those features of wavelets and multiresolution that are important for our presentation. There are many different ways of viewing and motivating wavelet expansions (see e.g. Daubechies, 1992). Our point of view in the present context is conveniently conveyed by the following example.

2.1 The Haar basis

The starting point is the box function $\phi(x) = \chi_{[0,1)}(x)$, which takes the value one on $[0, 1)$ and zero outside. The normalized dilates and translates $\phi_{j,k} = 2^{j/2} \phi(2^j \cdot - k)$, $k = 0, \dots, 2^j - 1$, of ϕ are readily seen to be *orthonormal*, that is, $(\phi_{j,k}, \phi_{j',k'})_{[0,1]} := \int_0^1 \phi_{j,k}(x) \phi_{j',k'}(x) dx = \delta_{j,k}$. Hence

$$P_j(f) := \sum_{k=0}^{2^j-1} \langle f, \phi_{j,k} \rangle \phi_{j,k}$$



is for each $j \in \mathbb{N}_0$ a simple orthogonal projector from $L_2([0, 1])$ onto the space S_j of piecewise constant functions subordinate to the dyadic mesh of size 2^{-j} . This projection resolves the function f up to scale j while finer details are averaged out; see Figure 1.

If the resolution is found to be insufficient, one has to discard previous efforts and recompute with a larger j . In contrast,

$$f = \sum_{j=0}^{\infty} (P_j - P_{j-1})f \quad (P_{-1} := 0) \quad (3)$$

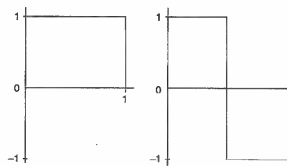
is a *multiscale representation* of the function f . Each term $(P_j - P_{j-1})f$ represents the detail in f at the scale 2^{-j} . Moreover, the layers of detail at each scale are mutually orthogonal.

As can be seen from Figure 2, to encode the difference information, one can use in place of the averaging profile $\phi(x)$ an oscillatory profile $\psi(x)$ – the *Haar wavelet* – given by

$$\psi(x) := \phi(2x) - \phi(2x - 1) \quad \text{which implies}$$

$$\phi(2x) = \frac{\phi(x) + \psi(x)}{2} \quad \text{and therefore}$$

$$\phi(2x - 1) = \frac{\phi(x) - \psi(x)}{2}$$



Hence, the fine scale averaging profiles can be recovered from a coarse scale average and an oscillatory profile. Thus, defining again $\phi_{j,k} := 2^{j/2} \phi(2^j \cdot - k)$, $\psi_{j,k} :=$

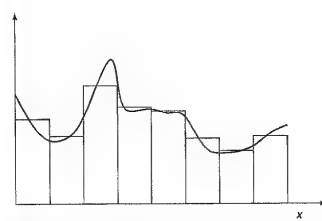


Figure 1. Piecewise constant approximation. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

$2^{j/2} \psi(2^j \cdot - k)$, one easily verifies the *two-scale relations*

$$\begin{aligned} \phi_{j,k} &= \frac{1}{\sqrt{2}} (\phi_{j+1,2k} + \phi_{j+1,2k+1}) \\ \psi_{j,k} &= \frac{1}{\sqrt{2}} (\phi_{j+1,2k} - \phi_{j+1,2k+1}) \\ \phi_{j+1,2k} &= \frac{1}{\sqrt{2}} (\phi_{j,k} + \psi_{j,k}) \\ \phi_{j+1,2k+1} &= \frac{1}{\sqrt{2}} (\phi_{j,k} - \psi_{j,k}) \end{aligned} \quad (4)$$

which give rise to a *change of basis*

$$\sum_{k=0}^{2^{j+1}-1} c_{j+1,k} \phi_{j+1,k} = \sum_{k=0}^{2^j-1} c_{j,k} \phi_{j,k} + \sum_{k=0}^{2^j-1} d_{j,k} \psi_{j,k} \quad (5)$$

where

$$\begin{aligned} c_{j,k} &= \frac{1}{\sqrt{2}} (c_{j+1,2k} + c_{j+1,2k+1}), \\ d_{j,k} &= \frac{1}{\sqrt{2}} (c_{j+1,2k} - c_{j+1,2k+1}) \\ c_{j+1,2k} &= \frac{1}{\sqrt{2}} (c_{j,k} + d_{j,k}), \end{aligned}$$

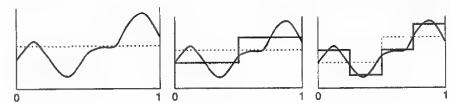


Figure 2. Different levels of resolution of f . A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

$$c_{j+1,2k+1} = \frac{1}{\sqrt{2}} (c_{j,k} - d_{j,k}) \quad (6)$$

Thus the representation in terms of fine scale averages can be obtained from the coarse scale averages already in hand by simply adding the detail (lost through the coarse projection) encoded in terms of the oscillatory profiles. The translated dilates $\psi_{j,k}(x) := 2^{j/2} \psi(2^j \cdot - k)$ are easily seen to be pairwise orthonormal

$$(\psi_{j,k}, \psi_{j',k'})_{\mathbb{R}} = \delta_{(j,k),(j',k')}, \quad j, j', k, k' \in \mathbb{Z} \quad (7)$$

Here and in the following, we use the notation $(f, g)_{\Omega} = \int_{\Omega} f g dx$ but suppress at times the subscript Ω when the reference to the domain is clear from the context.

Obviously, the above change of basis (5) can be repeated which gives rise to a cascading transform – the *fast wavelet transform*. It transforms a linear combination of fine scale box functions with an array of averages c_j into a linear combination of coarse scale box functions with coefficient array c_0 and Haar wavelets with arrays of detail coefficients d_j for each dyadic level $j < J$. This decomposition transform $T: c_j \rightarrow d' := (c_0, d_0, d_1, \dots, d_{J-1})$ looks schematically as follows:

$$\begin{array}{ccccccc} c_j & \rightarrow & c_{j-1} & \rightarrow & c_{j-2} & \rightarrow & \dots & \rightarrow & c_1 & \rightarrow & c_0 \\ & \searrow & & \searrow & & \searrow & & \searrow & & \searrow & \\ & d_{j-1} & & d_{j-2} & & \dots & & d_1 & & d_0 \end{array} \quad (8)$$

In other words, from c_j , we determine c_{j-1} and d_{j-1} by using (6), and so on. By (7) T is represented by a unitary matrix whose inverse is given by its transpose. Therefore the transform $T^{-1}: d' := (c_0, d_0, \dots, d_{J-1}) \rightarrow c_j$, which takes the detail coefficients into the single-scale average coefficients, has a similar structure that can also be read off from the relations (6):

$$\begin{array}{ccccccc} c_0 & \rightarrow & c_1 & \rightarrow & c_2 & \rightarrow & \dots & \rightarrow & c_{j-1} & \rightarrow & c_j \\ & \nearrow & & \nearrow & & \nearrow & & \nearrow & & \nearrow & \\ d_0 & & d_1 & & d_2 & & \dots & & d_{j-1} & & \end{array} \quad (9)$$

Thus, starting with c_0 and the wavelet coefficients d_0, \dots, d_{j-1} , we can use (9) to find c_j . Due to the cascading

structure and the fact that the relations in (6) involve only *finite filters or masks*, the number of operations required by both transforms is $O(2^j)$, i.e. stays proportional to the size of the arrays c_j .

In summary, there is a convenient and fast way of switching between different representations of a given projection $P_j(f) = \sum_k c_{j,k} \psi_{j,k}$, each having its advantages, as we shall see in subsequent discussions. From the theoretical point of view, since the wavelets allow us to encode the j th dyadic level of detail of a function f as

$$(P_{j+1} - P_j)f = \sum_{k=0}^{2^j-1} d_{j,k}(f) \psi_{j,k}, \quad d_{j,k}(f) := (f, \psi_{j,k})$$

the telescoping expansion (3) yields

$$f = P_0 f + \sum_{j=1}^{\infty} (P_j - P_{j-1})f = \sum_{j=1}^{\infty} \sum_{k=0}^{2^j-1} d_{j,k}(f) \psi_{j,k} =: d(f)^T \Psi \quad (10)$$

The convergence of this expansion in L_2 follows from the convergence of the orthogonal projections in L_2 .

A wavelet decomposition of a function $f \in L_2$ is analogous in spirit to the decimal representations of a real number. The wavelet coefficients play the role of *digits*; receiving more wavelet coefficients gives us progressively better accuracy in representing f . Of course, the classical Fourier transform of periodic functions and also Taylor expansions do, in principle, the same. The particular advantages of wavelet representations rely to a large extent on the following facts. First of all, the orthonormality of the $\psi_{j,k}$ gives

$$\|f\|_{L_2} = \left(\sum_{j=0}^{\infty} \|(P_j - P_{j-1})f\|_{L_2}^2 \right)^{1/2} = \|d(f)\|_{\ell_2} \quad (11)$$

That is, there is a tight relation between the function and coefficient norm. Thus perturbing the digits, which will happen in every computation, in particular, discarding small digits, will change the function norm only by the same small amount. Clearly, the convergence of the series implies that the digits will eventually have to become arbitrarily small. However, which digits become how small can easily be inferred from *local properties* of f . In fact, since the $\psi_{j,k}$ are orthogonal to constants – they have *first order vanishing moments* – one has for $S_{j,k} := \text{supp } \psi_{j,k} = 2^{-j}[k, k+1]$

$$|d_{j,k}(f)| = \inf_{c \in \mathbb{R}} |(f - c, \psi_{j,k})| \leq \inf_{c \in \mathbb{R}} \|f - c\|_{L_2(S_{j,k})} \leq 2^{-j} \|f'\|_{L_2(S_{j,k})} \quad (12)$$

where the last estimate follows, for example, from Taylor's expansion. Thus, $d_{j,k}(f)$ is small when $f|_{S_{j,k}}$ is smooth.

2.2 Biorthogonal wavelets on \mathbb{R}

The Haar basis is, of course, not suitable when, for instance, higher regularity of the approximation system is required. The discovery by I. Daubechies (Daubechies, 1992), of a family of compactly supported orthonormal wavelets in $L_2(\mathbb{R})$ of arbitrary high regularity opened the door to a wide range of applications. Of perhaps even more practical relevance was the subsequent construction of *biorthogonal wavelets* put forward by Cohen, Daubechies and Feauveau (1992). The biorthogonal approach sacrifices L_2 -orthogonality in favor of other properties such as symmetry of the basis functions and better localization of their supports.

The construction of biorthogonal wavelets starts with a *dual pair* of compactly supported scaling functions $\phi, \tilde{\phi}$, that is,

$$(\phi, \tilde{\phi}(\cdot - k)) = \delta_{0,k}, \quad k \in \mathbb{Z} \quad (13)$$

that satisfy the two scale relations

$$\phi(x) = \sum_{k \in \mathbb{Z}} a_k \phi(2x - k), \quad \tilde{\phi}(x) = \sum_{k \in \mathbb{Z}} \tilde{a}_k \tilde{\phi}(2x - k) \quad (14)$$

with finitely supported *masks* $(a_k)_{k \in \mathbb{Z}}, (\tilde{a}_k)_{k \in \mathbb{Z}}$; see (4). Each of the functions

$$\begin{aligned} \psi(x) &= \sum_{k \in \mathbb{Z}} (-1)^k \tilde{a}_{1-k} \phi(2x - k) \\ \tilde{\psi}(x) &= \sum_{k \in \mathbb{Z}} (-1)^k a_{1-k} \tilde{\phi}(2x - k) \end{aligned} \quad (15)$$

generates by means of shifts and dilates a *biorthogonal basis* for $L_2(\mathbb{R})$. Each $f \in L_2(\mathbb{R})$ has the following unique expansions:

$$f = \sum_{j=-1}^{\infty} \sum_{k \in \mathbb{Z}} (f, \tilde{\psi}_{j,k})_{\mathbb{R}} \psi_{j,k} = \sum_{j=-1}^{\infty} \sum_{k \in \mathbb{Z}} (f, \psi_{j,k})_{\mathbb{R}} \tilde{\psi}_{j,k} \quad (16)$$

where we have used the notation, $\psi_{-1,k} := \phi_{0,k}$, $\tilde{\psi}_{-1,k} := \tilde{\phi}_{0,k}$. One thus has $(\psi_{j,k}, \tilde{\psi}_{l,m})_{\mathbb{R}} = \delta_{(j,l), (k,m)}$.

Each of these systems is a *Riesz basis*, which means [1]

$$\|f\|_{L_2(\mathbb{R})}^2 \sim \sum_{j=-1}^{\infty} \sum_{k \in \mathbb{Z}} |(f, \tilde{\psi}_{j,k})|^2 \sim \sum_{j=-1}^{\infty} \sum_{k \in \mathbb{Z}} |(f, \psi_{j,k})|^2 \quad (17)$$

The inequalities (17) ensure a tight relation between the function norm and the coefficient norm.

Cohen, Daubechies and Feauveau (1992) construct a family of biorthogonal pairs with each of $\psi, \tilde{\psi}$ of compact support. Given any desired order r of differentiability, one can find a biorthogonal pair in this family with ψ having r continuous derivatives. Moreover, one can also require that a suitable linear combination of $(\phi(\cdot - k))_{k \in \mathbb{Z}}$ (respectively $(\tilde{\phi}(\cdot - k))_{k \in \mathbb{Z}}$) will represent any given polynomial of order $\leq m$, (respectively \tilde{m}). The biorthogonality relations then imply that the wavelets $\psi_{j,k}, \tilde{\psi}_{j,k}$ (for $j \geq 0$) are orthogonal to all polynomials of order \tilde{m}, m respectively. An analogous argument to (12) then shows that the coefficients $(f, \tilde{\psi}_{j,k}), (f, \psi_{j,k})$ decay like $2^{-mj}, 2^{-\tilde{m}j}$ when f has bounded derivatives on the supports of $\tilde{\psi}_{j,k}, \psi_{j,k}$ of order m, \tilde{m} , respectively, in L_2 . Thus higher local smoothness results in a stronger size reduction of corresponding wavelet coefficients.

The setting of biorthogonal wavelets is particularly appealing from a practical point of view since the *primal generator* ϕ can be chosen as any B-spline, and in turn the *primal wavelet generator* ψ is also a spline function with an explicit – piecewise polynomial – analytical expression.

2.3 Wavelets on domains

Biorthogonal wavelets provide a beautiful and conceptually simple multiscale decomposition of functions. They offer great versatility in the choice of the basis and dual elements including compact support, smoothness, and even piecewise polynomial structure. Meanwhile they maintain the essential features of orthogonal wavelet decompositions such as norm equivalences and cancellation properties. Moreover, in connection with differential equations, the space L_2 often plays only a secondary or artificial role. Therefore, dispensing with L_2 -orthogonality is, in general, not even a quantitative drawback.

That is the good news. The bad news is that the above wavelet constructions are inherently made on \mathbb{R}, \mathbb{R}^d , or a torus. In numerical applications, the setting is typically on a finite domain or manifold Ω . Such domains and manifolds do not maintain the dilation and translation structure of the full Euclidean space or the torus. [2] Fortunately, there are constructions of multiscale bases tailored to general domains and manifolds. Albeit, these come at some expense of a certain level of technicality. In order not to destroy the main flow of this chapter, we shall only give an overview of some of the ideas used for these constructions. The reader can consult Dahmen (1997) and the references quoted there for a more detailed description of the construction of multiscale bases on (bounded) domains and manifolds.

The starting point of these constructions is again *multiresolution*. By this, we mean a hierarchy of (now finite

dimensional) subspaces S_j of some function space \mathcal{X}

$$S_0 \subset S_1 \subset S_2 \subset \dots \subset \mathcal{X}, \quad \bigcup_j S_j = \mathcal{X}$$

that are spanned by *single-scale bases* $S_j = \text{span } \Phi_j := S(\Phi_j)$, $\Phi_j = \{\phi_\gamma; \gamma \in \mathcal{I}_j\}$. The space \mathcal{X} is typically an L_p or Sobolev space. It is important that the bases Φ_j are *scalewise stable* with respect to some discrete norm $\|\cdot\|$ in the sense that

$$\|(\|c_\gamma \phi_\gamma\|_{\mathcal{X}})_{\gamma \in \mathcal{I}_j}\| \sim \|\sum_{\gamma \in \mathcal{I}_j} c_\gamma \phi_\gamma\|_{\mathcal{X}} \quad (18)$$

with constants that do not depend on the level j . In the case where $\mathcal{X} = L_p$ or W_p^r , the discrete norm $\|\cdot\|$ is typically the ℓ_p norm. For instance, Φ_j could be a finite element nodal basis on a j -fold refinement of some initial mesh for Ω . In this example, the indices γ may represent the vertices in the mesh. One then looks for *decompositions*

$$S_{j+1} = S_j \oplus W_j, \quad W_j = \text{span } \Psi_j, \quad \Psi_j = \{\psi_\lambda; \lambda \in \mathcal{J}_j\}$$

The *multiscale basis*

$$\Psi := \bigcup_{j=-1}^{\infty} \Psi_j =: \{\psi_\lambda; \lambda \in \mathcal{J}\} \quad (\Psi_{-1} := \Phi_0)$$

is then a candidate for a wavelet basis. At this point a word on notation is in order. The index set has two component subsets: $\mathcal{J} = \mathcal{J}_0 \cup \mathcal{J}_\Psi$. The index set \mathcal{J}_0 has a finite cardinality and labels the basis functions in S_0 of 'scaling function' type. The true wavelets correspond to the indices in \mathcal{J}_Ψ . These indices absorb different types of information such as the scale $j = |\lambda|$, the spatial location $k(\lambda)$, or, when dealing with a spatial dimension $d > 1$, the type $\epsilon(\lambda)$ of ψ_λ . An example is $\psi_\lambda(x, y) = 2^{j/2} \psi(2^j(x, y) - (k, l)) = 2^{j/2} \psi(2^j x - k) 2^{l/2} \tilde{\psi}(2^j y - l)$ where $\lambda \leftrightarrow (j, (k, l), (1, 0))$.

Of course, there is a continuum of possible complements W_j and the question arises as to what are 'good complements'. The previous section already indicates the role of biorthogonality in this context. So a typical strategy is to split the multiresolution spaces S_j in such a way that there exists a biorthogonal or dual collection $\tilde{\Psi}$, corresponding to a dual *multiresolution sequence* (\tilde{S}_j) , that belongs to the dual \mathcal{X}' in such a way that

$$(\psi_\lambda, \tilde{\psi}_\mu) = \delta_{\lambda, \mu}, \quad \lambda, \mu \in \mathcal{J}$$

and hence

$$f = \sum_{\lambda \in \mathcal{J}} (f, \tilde{\psi}_\lambda) \psi_\lambda \quad (19)$$

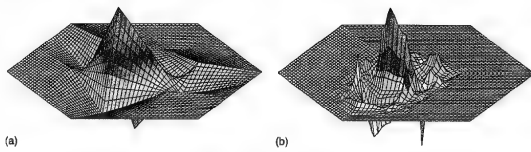


Figure 3. (a) Primal wavelet (b) Dual wavelet. A color version of this image is available at <http://www.mrw.interscience.wiley.com/cem>

The classical situation is $\mathcal{X} = \mathcal{X}' = L_2(\Omega)$ so that one has in this case the alternate representation $f = \sum_{\lambda \in \Lambda} \langle f, \psi_\lambda \rangle \psi_\lambda$ (see Carnicer, Dahmen and Peña, 1996; Cohen, Daubechies and Peauveau, 1992; Dahmen, 1994, 1996).

When no global smoothness is required, the concept of *multiwavelets* (Alpert, 1993) offers a convenient way of generalizing the Haar basis to higher order accuracy and cancellation properties; see, for example, von Petersdorff, Schneider and Schwab (1997) for an application to second kind integral equations.

We describe now one concrete approach (see Canuto, Tabacco and Urban, 1999, 2000; Cohen and Masson, 1997; Dahmen and Schneider, 1999a,b) based on domain decomposition that works for $\mathcal{X} = L_2(\Omega)$ and realizes at least global continuity. Suppose that

$$\Omega = \bigcup_{1 \leq m \leq M} \kappa_m(\square), \quad \kappa_m: \square \rightarrow \Omega_m$$

where each κ_i is a regular mapping from a parameter domain such as the unit d -cube, into a subdomain Ω_i of Ω . A wavelet basis Ψ^Ω is then constructed along the following lines:

- Start with biorthogonal wavelet bases $\Psi^R, \tilde{\Psi}^R$ on \mathbb{R} and adapt them to biorthogonal wavelet bases $\Psi^I, \tilde{\Psi}^I$ on $I = [0, 1]$.
- Use tensor products to obtain bases Ψ^\square on the unit cube $\square = [0, 1]^d$.
- Use parametric liftings to derive bases $\Psi^\Omega = \Psi^\square \circ \kappa_i^{-1}$ on $\Omega_i = \kappa_m(\square)$, which then have to be glued together to produce, for example, globally continuous bases $\Psi = \Psi^\Omega$ on Ω (see e.g. Canuto, Tabacco and Urban, 1999, 2000; Cohen and Masson, 1997; Dahmen and Schneider, 1999a). An alternative approach leads to wavelets of arbitrary regularity permitted by the regularity of the domain (Dahmen and Schneider, 1999b).

The following Figure 3 has been provided as a courtesy by H. Harbrecht. It displays an example of a globally continuous primal and dual wavelet on a two dimensional patchwise defined manifold in which the supports cross the patch boundaries.

Alternatively, hierarchies of uniform refinements of an arbitrary initial triangulation can be used to construct finite element based wavelets (see e.g. Dahmen and Stevenson, 1999; Stevenson, 2000).

All these constructions aim to realize (19). However, biorthogonality by itself is not quite sufficient in general to guarantee relations like (17). It is shown in Dahmen (1996) that if in addition the multiresolution spaces S_j and \tilde{S}_j each satisfy a (quite general form of a) direct (sometimes referred to as a *Jackson*) estimate which quantifies their approximation properties, and in addition satisfy an inverse (referred to as *Bernstein*) estimate, which quantifies the regularity of these spaces, then an $L_2(\Omega)$ -norm equivalences of the form (17) hold. One actually obtains norm equivalences for a whole range of smoothness spaces, (possibly weighted) Sobolev spaces, around L_2 ; a fact that is actually more important for the intended applications (Dahmen, 1996; Dahmen and Stevenson, 1999).

The above approach, in particular (Dahmen and Stevenson, 1999), can be viewed as a special realization of the following general strategy. To describe this approach, it is now convenient to view a (countable) collection Θ of functions, such as a wavelet basis or a basis of scaling functions, as a column vector based on some fixed but unspecified ordering of its elements. Refinement relations of the form (14) take then the form $\Phi_j^T = \Phi_{j+1}^T M_{j,0}$ where the columns of the matrix $M_{j,0}$ consist of the mask coefficients in each two-scale relation for the elements of the scaling functions on level j . For instance, in the case of the Haar basis on $[0, 1]$, (4) says that each column in $M_{j,0}$ has at two successive positions the value $2^{-1/2}$ as the only nonzero entry. This format persists in much wider generality and can be used to represent two-scale relations for any hierarchy S of nested spaces spanned by scaling function bases $S_j = \text{span } \Phi_j$. In the same way, a basis Ψ_j

spanning some complement W_j of S_j in S_{j+1} , has the form $\Psi_j^T = \Phi_{j+1}^T M_{j,1}$. It is easy to see that $M_{j,1}$ completes $M_{j,0}$ to an invertible operator $M_j := (M_{j,0}, M_{j,1})$ if and only if $S_{j+1} = S_j \oplus W_j$ and that the complement bases are *uniformly scalewise stable* in the sense of (18) if and only if the condition numbers of the M_j with respect to the corresponding norms are uniformly bounded (Carnicer, Dahmen and Peña, 1996). Of course, in the case of orthonormal bases one has $G_j = M_j^T$.

One can now define multiscale transformations that change, for instance, the representation of an element in S_j with respect to Φ_j into the representation with respect to Φ_0 and the complement bases Ψ_j , $j < J$, in complete analogy to (8) and (9). In fact, the refinement relations imply that

$$e^{j+1} = (M_{j,0} e^j + M_{j,1} d^j) \quad (20)$$

The refinement matrix $M_{j,0}$ can be viewed as a *prediction operator*. When the detail coefficients are zero, it provides an exact representation of the data on the next higher level of resolution. It is also sometimes called a *subdivision operator* (see e.g. Arandiga, Donat and Harten, 1998, 1999; Carnicer, Dahmen and Peña, 1996). It follows from (20) that the transformation T_j^{-1} , taking the detail coefficients into the single-scale coefficients e^j , uses the matrices M_j in each cascading step of (9).

Conversely, setting $G_j := M_j^{-1} = \begin{pmatrix} G_{j,0} \\ G_{j,1} \end{pmatrix}$, one has

$$e^j = G_{j,0} e^{j+1} + d^j = G_{j,1} e^{j+1} \quad (21)$$

Hence the transformation T_j that decomposes e^j into details d^j and coarse scale coefficients e^j has the same cascading structure as (8), now based on the filter matrices G_j .

Furthermore, one can show that the transformations T_j have uniformly bounded spectral condition numbers independent of J if and only if the corresponding union Ψ of the complement bases Ψ_j and the coarse scale basis Φ_0 forms a Riesz basis for L_2 (Dahmen, 1994; Carnicer, Dahmen and Peña, 1996).

While it is often difficult to directly construct a Riesz basis for the space of interest, in many cases, it is easy to find for each level j some *initial* complement bases Ψ_j . For instance, when working with a hierarchy of nodal finite element bases, complement bases are provided by the *hierarchical basis* consisting of those nodal basis functions at the nodes of the next higher level of resolution (see e.g. Yserentant, 1986). As a second step one can then generate from this initial multiscale decomposition another one that has certain desirable properties, for instance, a higher order of vanishing moments. The important point to be made in this regard is that all of

this can be done completely on a discrete level. To this end, suppose that an *initial* completion $M_{j,1}$ of the refinement matrix $M_{j,0}$ (and G_j) is known. Then all other *stable* completions have the form (Carnicer, Dahmen and Peña, 1996)

$$M_{j,1} = M_{j,0} L + \tilde{M}_{j,1} K \quad (22)$$

with inverse blocks

$$G_{j,0} = \tilde{G}_{j,0} - \tilde{G}_{j,1} (K^T)^{-1} L^T, \quad G_{j,1} = \tilde{G}_{j,1} (K^T)^{-1} \quad (23)$$

In fact, this follows from the identity

$$I = \tilde{M}_j \tilde{G}_j = \tilde{M}_j \begin{pmatrix} I & L \\ 0 & K \end{pmatrix} \begin{pmatrix} I & -LK^{-1} \\ 0 & K^{-1} \end{pmatrix} \tilde{G}_j =: M_j G_j$$

The special case $K = I$ is often referred to as *Lifting Scheme*, (Sweldens, 1996, 1998). The parameters in the matrices L, K can be used to modify the complement bases. Such modifications of stable completions are used for instance in the construction of wavelets on an interval (see e.g. Dahmen, Kunoth and Urban, 1999) and hence in the above-mentioned domain decomposition approach (Canuto, Tabacco and Urban, 1999, 2000; Cohen and Masson, 1997; Dahmen and Schneider, 1999a), as well as in the construction of finite element based wavelets through coarse grid corrections (Carnicer, Dahmen and Peña, 1996; Dahmen, 1997; Dahmen and Kunoth, 1992; Dahmen and Stevenson, 1999; Stevenson, 2000; Vassilevski and Wang, 1997). A further important application concerns raising the order of *vanishing moments*: Choose $K = I$ and L such that

$$\int_{\Omega} \Psi_j^T P \, dx = \int_{\Omega} \Phi_{j+1}^T M_{j,1} P \, dx = \int_{\Omega} \Phi_j^T L P \, dx + \tilde{\Psi}_j^T P \, dx = 0, \quad P \in \mathcal{P}_m \quad (24)$$

The significance of the above formulations lies in the versatility of handling multiscale decompositions entirely on a discrete level. This allows one to circumvent (at least to some extent) the explicit construction of complicated basis functions (see Harten, 1996). However, statements about stability are often based on explicit knowledge of the underlying multiscale bases.

2.4 The key features

The ideas put forward in the previous section allow one to construct multiscale bases for a variety of domains and

even closed surfaces. In this section, we collect the main properties of these constructions that are valid on a domain Ω of spatial dimension d . We shall assume these properties in our subsequent applications. These key properties can be summarized as follows:

- **Locality (L)** • **Cancellation Properties (CP)**
- **Norm Equivalences (NE).**

Locality: (L) means that the elements of Ψ all have compact support $S_\lambda := \text{supp } \psi_\lambda$ that scales properly, that is,

$$\text{diam}(S_\lambda) \sim 2^{-|\lambda|} \quad (25)$$

Locality is crucial for applications on bounded domains and for the efficiency of associated multiscale transforms.

Cancellation Properties: (CP) generalizes our earlier observation (12). It means that integrating a wavelet against a locally smooth function acts like differencing. Assume for example that the wavelets are normalized in L_2 , that is $\|\psi_\lambda\|_{L_2} \sim 1$. Cancellation will then mean that

$$|\langle v, \psi_\lambda \rangle| \lesssim 2^{-|\lambda|(\tilde{m} + (d/2) - (d/p))} \|v\|_{W_p^{\tilde{m}}(S_\lambda)}, \quad \lambda \in \mathcal{J}_\Psi \quad (26)$$

where $\|v\|_{W_p^{\tilde{m}}(S_\lambda)}$ is the usual n th order seminorm of the corresponding Sobolev space on the domain G . Analogous relations can of course be formulated for the dual basis $\tilde{\Psi}$.

The integer \tilde{m} signifies the strength of the cancellation properties because it says up to which order the local smoothness of the function is rewarded by the smallness of the coefficients (in this case of the dual expansion). Obviously, when Ω is a Euclidean domain, (26) implies that the wavelets have *vanishing polynomial moments* of order \tilde{m} , that is,

$$\langle P, \psi_\lambda \rangle_\Omega = 0, \quad P \in \mathbb{P}_\beta, \quad \lambda \in \mathcal{J}_\Psi \quad (27)$$

Conversely, as in (12), the vanishing moments imply that for $[(1/p) + (1/p')] = 1$

$$\begin{aligned} |\langle v, \psi_\lambda \rangle| &= \inf_{P \in \mathbb{P}_\beta} |\langle v - P, \psi_\lambda \rangle| \\ &\leq \inf_{P \in \mathbb{P}_\beta} \|v - P\|_{L_p(S_\lambda)} \|\psi_\lambda\|_{L_{p'}}, \\ &\lesssim 2^{-|\lambda|((d/2) - (d/p))} \inf_{P \in \mathbb{P}_\beta} \|v - P\|_{L_p(S_\lambda)} \end{aligned}$$

where we have used that

$$\begin{aligned} \|\psi_\lambda\|_{L_p} &\sim 2^{|\lambda|(d/p) - (d/2)} \sim 2^{|\lambda|((d/2) - (d/p))} \quad \text{when} \\ \|\psi_\lambda\|_{L_2} &\sim 1 \end{aligned} \quad (28)$$

Now standard estimates on *local polynomial approximation* (see e.g. DeVore and Sharpley, 1984) tell us that

$$\inf_{P \in \mathbb{P}_\beta} \|v - P\|_{L_p(G)} \lesssim (\text{diam } G)^k \|v\|_{W_p^k(G)}$$

which yields (26). We refer to (26) as the *cancellation property* of order \tilde{m} rather than to (27), since it makes sense for domains where ordinary polynomials are not defined.

Norm Equivalences: The cancellation properties tell us under what circumstances wavelet coefficients are small. One expects to have only relatively few significant coefficients when the expanded function is very smooth except for singularities on lower dimensional manifolds. This helps to recover a function with possibly few coefficients only if small perturbations in the coefficients give rise to perturbations of the function that are also small with respect to the relevant norm. Recall that for function spaces \mathcal{X} with local norms, it is usually easy to construct multiscale bases Ψ that are uniformly *scalewise* stable, that is,

$$\left\| \sum_{\lambda=j} d_\lambda \psi_\lambda \right\|_{\mathcal{X}} \sim \|(\|d_\lambda \psi_\lambda\|_{\mathcal{X}})_{\lambda=j}\| \quad (29)$$

uniformly in j , where $\|\cdot\|$ is some appropriate discrete norm.

In some cases, this stability property can be extended to the whole array Ψ over all scales. In the particular case when $\mathcal{X} = \mathcal{H}$ is a Hilbert space, this is expressed by saying that, with the normalization $\|\psi_\lambda\|_{\mathcal{H}} \sim 1$, the family Ψ is a *Riesz basis* for the whole function space \mathcal{H} , that is, every element $v \in \mathcal{H}$ possesses a unique expansion in terms of Ψ and there exist finite positive constants c_Ψ, C_Ψ such that

$$\begin{aligned} c_\Psi \|(v_\lambda)_\lambda\|_{\ell_2} &\leq \left\| \sum_\lambda v_\lambda \psi_\lambda \right\|_{\mathcal{H}} \leq C_\Psi \|(v_\lambda)_\lambda\|_{\ell_2}, \\ \forall v &= (v_\lambda)_\lambda \in \ell_2 \end{aligned} \quad (30)$$

Thus, while relaxing the requirement of orthonormality, a Riesz basis still establishes a strong coupling between the continuous world, in which the mathematical model is often formulated, and the discrete realm which is more apt to computational realizations. Therefore, it should not be a surprise that the availability of such bases for function spaces may be exploited for numerical methods.

We shall exploit norm equivalences for the problem class (2) where the relevant spaces are *Sobolev spaces* or tensor products of them. Recall that for $n \in \mathbb{N}$ the space $H^n(\Omega)$ consists of those elements of $L_2(\Omega)$ whose n th order weak derivatives are also in $L_2(\Omega)$. More generally,

for $1 \leq p \leq \infty$, we have

$$W_p^n(\Omega) := \{f: \partial^\alpha f \in L_p(\Omega), |\alpha| \leq n\} \quad (31)$$

and the corresponding (semi-)norms are given by $\|f\|_{W_p^n(\Omega)} := (\sum_{|\alpha|=n} \|\partial^\alpha f\|_{L_p(\Omega)})^{1/p}$ and $\|v\|_{W_p^n(\Omega)} := (\sum_{m=0}^n \|f\|_{W_p^m(\Omega)})^{1/p}$. Dealing with traces of functions on boundary manifolds, for instance, forces one to consider also non-integer smoothness orders $t \in \mathbb{R}$. For $t > 0$, these spaces can be defined either by interpolation between spaces of integer order (see e.g. Bergh and Löfström, 1976) or directly through intrinsic norms of the form

$$\begin{aligned} \|v\|_{W_p^t(\Omega)} &= \left(\|v\|_{W_p^0(\Omega)}^p + \sum_{|\alpha|=n} \int_\Omega \int_\Omega \frac{|\partial^\alpha v(x) - \partial^\alpha v(y)|^p}{|x - y|^{d+(t-n)p}} dx dy \right)^{1/p} \\ n &:= [t] \end{aligned}$$

Moreover, for Lipschitz domains Ω and $\Gamma \subset \partial\Omega$, we denote by $H_{0,\Gamma}^s(\Omega)$ the closure of those C^∞ functions on Ω with respect to the H^s -norm that vanish on Γ . We briefly write $H_0^s(\Omega)$ when $\Gamma = \partial\Omega$. We refer to Adams (1978) for more details on Sobolev spaces.

In the sequel for $t \geq 0$, H^t will denote some closed subspace of $H^t(\Omega)$ either of the form $H_0^t(\Omega) \subseteq H^t \subseteq H^t(\Omega)$ or with finite codimension in $H^t(\Omega)$. For $t < 0$, we define H^t as the *dual space* $H^t = (H^{-t})'$. Starting with a suitable wavelet Riesz basis Ψ for $\mathcal{H} = L_2(\Omega)$, a whole family of realizations of (30) can be formulated as follows. There exist positive constants $\gamma, \tilde{\gamma} > 0$ (depending on the regularity of the wavelet basis) with the following property: For $s \in (-\tilde{\gamma}, \gamma)$, there exist positive constants c_s, C_s such that every $v \in H^s$ possesses a unique expansion $v = \sum_{\lambda \in \mathcal{J}} v_\lambda 2^{-|\lambda|} \psi_\lambda$ such that

$$c_s \|(v_\lambda)_\lambda\|_{\ell_2} \leq \left\| \sum_{\lambda \in \mathcal{J}} v_\lambda 2^{-|\lambda|} \psi_\lambda \right\|_{H^s} \leq C_s \|(v_\lambda)_\lambda\|_{\ell_2} \quad (32)$$

Thus properly scaled versions of the wavelet basis Ψ for L_2 are Riesz bases for a *whole range* of smoothness spaces, including of course $s = 0$ as a special case. This range depends on the *regularity* of the wavelets. In many constructions, one has $\gamma = 3/2$ corresponding to globally continuous wavelets (Canuto, Tabacco and Urban, 1999, 2000, Cohen and Masson, 1997; Dahmen and Schneider, 1999a; Dahmen and Stevenson, 1999).

Establishing (32) works actually the other way around. It is usually easier to verify (32) for *positive* s . This can be derived from the validity of *Bernstein and Jackson* estimates for the *primal* multiresolution sequences only. If one can do this, however, for the *primal* and for the

dual multiresolution sequences associated with a dual pair of multiscale bases $\Psi, \tilde{\Psi}$, (32) follows for the whole range of regularity indices s by an interpolation argument (see e.g. Dahmen, 1996; Dahmen and Stevenson, 1999). In particular, this says that the Riesz basis property for $L_2(\Omega)$ follows from that of scaled versions of $\Psi, \tilde{\Psi}$ for positive Sobolev regularity (see also Cohen, 2000, 2003; Dahmen, 1997, 2003).

Remark 1. We emphasize the case (32) because it implies further relations that will be important later for robustness. To describe these, recall our convention of viewing a collection Θ of basis functions sometimes as a vector whose entries are ordered in a fixed but unspecified way. Ordering the wavelet coefficient arrays in a natural way, we can write $\sum_{\lambda \in \mathcal{J}} v_\lambda 2^{-|\lambda|} \psi_\lambda =: \mathbf{v}^T \mathbf{D}^s \Psi$ where $(\mathbf{D}^s)_{\lambda, \mu} := (2^{|\lambda|} \delta_{\lambda, \mu})_{\lambda, \mu}$ and $\mathbf{v} := (v_\lambda)_{\lambda \in \mathcal{J}}$. In the problem class (2), one often encounters Hilbert (energy) spaces endowed with a norm of the type $\|v\|_{\mathcal{H}}^2 := \epsilon(\nabla v, \nabla v) + (v, v)$. The performance of multilevel preconditioners for such problems often depends on ϵ . It will be seen that a remedy for this can be based on *robust* equivalences of the following form that can be derived from (32) for $s = 0$ and $s = 1$ (Cohen, Dahmen and DeVore, 2001; Dahmen, 2001): assume that $\gamma > 1$ and define the diagonal matrix $\mathbf{D}_\epsilon := ((1 + \sqrt{\epsilon}) 2^{|\lambda|})_{\lambda, \mu} \delta_{\lambda, \mu} \mathbf{e}_{\mathcal{J}}$. Then

$$\begin{aligned} (2(c_0^{-2} + c_1^{-2}))^{-1/2} \|\mathbf{v}^T \mathbf{D}_\epsilon^{-1} \Psi\|_{\mathcal{H}} &\leq \|\mathbf{v}^T \mathbf{D}_\epsilon^{-1} \Psi\|_{\mathcal{H}} \\ &\leq (C_0^{-2} + C_1^{-2})^{1/2} \|\mathbf{v}\|_{\ell_2} \end{aligned} \quad (33)$$

We wish to conclude this section with the following remarks concerning *duality*; see, for example, Dahmen (2003) for more details. As indicated before, the known constructions of a wavelet basis Ψ , that satisfy norm equivalences of the form (32), involve to some extent the simultaneous construction of a dual basis $\tilde{\Psi}$. Conversely, the existence of such a dual basis is actually a consequence of the Riesz basis property in the following sense. It is not hard to show that the validity of (30) implies the existence of a collection $\tilde{\Psi} \subset \mathcal{H}'$ such that $\langle \psi_\lambda, \tilde{\psi}_\mu \rangle = \delta_{\lambda, \mu}$, where $\langle \cdot, \cdot \rangle$ is the duality pairing that identifies the representation of \mathcal{H}' . Moreover, $\tilde{\Psi}$ is a Riesz basis for \mathcal{H}' , that is,

$$C_{\tilde{\Psi}}^{-1} \|\mathbf{w}\|_{\ell_2} \leq \|\mathbf{w}^T \tilde{\Psi}\|_{\mathcal{H}'} \leq C_{\tilde{\Psi}} \|\mathbf{w}\|_{\ell_2}, \quad \mathbf{w} \in \ell_2 \quad (34)$$

This will mainly be used in the equivalent form

$$C_{\tilde{\Psi}}^{-1} \langle \Psi, v \rangle_{\mathcal{H}'} \|v\|_{\ell_2} \leq \|\mathbf{v}\|_{\mathcal{H}'} \leq C_{\tilde{\Psi}} \langle \Psi, v \rangle_{\ell_2} \quad (35)$$

where we have abbreviated $\langle \Psi, v \rangle := (\langle \psi_\lambda, v \rangle : \lambda \in \mathcal{J})^T$.

In particular, if we want to construct a Riesz basis for L_2 , then the dual basis (with respect to the L_2 -inner product as a dual pairing) must also be a Riesz basis in L_2 , in agreement with the above remarks concerning (32). This rules out the practically convenient so-called *hierarchical bases* induced by interpolatory scaling functions since the dual basis is essentially comprised of Dirac distributions. An important further example is $\mathcal{H} = H_0^1(\Omega)$ for which, according to (32), a Riesz basis is obtained by renormalizing the functions ψ_λ with the weight $2^{-|\lambda|}$, which also amounts to redefining Ψ as $D^{-1}\Psi$. In this case, (35) offers a convenient way of evaluating the H^{-1} -norm which is useful for least squares formulations of second order elliptic problems or their mixed formulations (Dahmen, Kunoth and Schneider, 2002). Note that in this particular case, unlike the situation in Remark 1, Ψ need not be a Riesz basis for L_2 .

2.5 Wavelets and linear operators

So far we have focused on wavelet representations of functions. For the problem class (2), in particular, it will be important to deal with wavelet representations of operators. In this section, we collect a few important facts concerning linear operators that follow from the above features. As a simple guiding example consider Poisson's equation on some bounded domain $\Omega \subset \mathbb{R}^d$

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \Gamma := \partial\Omega \quad (36)$$

It will be crucial to interpret this equation properly. Multiplying both sides of (36) by smooth test functions that vanish on Γ , and integrating by parts, shows that the solution u satisfies

$$(\nabla v, \nabla u) = (v, f) \quad \text{for all smooth } v \quad (37)$$

where $(v, w) := \int_\Omega v w \, dx$. However, this latter form makes sense even when u belongs only to the Sobolev space $H_0^1(\Omega)$ (recall Section 2.4) and when the test functions also just belong to $H_0^1(\Omega)$. Moreover, the right hand side makes sense whenever f is only a distribution in the dual $H^{-1}(\Omega)$ of $H_0^1(\Omega)$. Here (\cdot, \cdot) is then understood to be the dual form on $H_0^1(\Omega) \times H^{-1}(\Omega)$ induced by the standard L_2 -inner product. Thus, defining the linear operator \mathcal{A} by $(\nabla v, \nabla u) = (v, \mathcal{A}u)$ for all $v, u \in H_0^1(\Omega)$, the boundary value problem (36) is equivalent to the operator equation

$$\mathcal{A}u = f \quad (38)$$

where, roughly speaking, \mathcal{A} is in this case the Laplacian (with incorporated homogeneous boundary conditions), taking $H_0^1(\Omega)$ into its dual $H^{-1}(\Omega)$.

The Standard Wavelet Representation: Suppose now that as in the Laplace case described above, we have a linear operator $\mathcal{A} : \mathcal{H} \rightarrow \mathcal{H}'$ and that Ψ is a Riesz-basis for \mathcal{H} , that is, (30) holds. Then for any $v = \sum_\lambda v_\lambda \psi_\lambda \in \mathcal{H}$ one has

$$\begin{aligned} \mathcal{A}v &= \sum_\lambda (\psi_\lambda, \mathcal{A}v) \tilde{\psi}_\lambda = \sum_\lambda \left(\psi_\lambda, \mathcal{A} \left(\sum_\nu v_\nu \psi_\nu \right) \right) \tilde{\psi}_\lambda \\ &= \sum_\lambda \left(\sum_\nu (\psi_\lambda, \mathcal{A} \psi_\nu) v_\nu \right) \tilde{\psi}_\lambda \end{aligned}$$

Thus the coefficient array w of $\mathcal{A}v \in \mathcal{H}'$ with respect to the dual basis $\tilde{\Psi}$ is given by

$$w = Av \quad \text{where} \quad A := ((\psi_\lambda, \mathcal{A} \psi_\nu))_{\lambda, \nu}, \quad v = (v_\nu)_\nu \quad (39)$$

The above example (36) is a typical application where Ψ and $\tilde{\Psi}$ are biorthogonal Riesz bases in the Sobolev space $\mathcal{H} = H_0^1(\Omega)$ and its dual $\mathcal{H}' = H^{-1}(\Omega)$ respectively.

A is often referred to as the *standard wavelet representation* of \mathcal{A} . Note that in conventional discretizations such as finite elements and finite differences, the operators can usually only be *approximated*. A basis allows one to capture, at least conceptually, all of the full infinite dimensional operator, a fact that will later be seen to have important consequences.

Well-Posedness and an Equivalent ℓ_2 -Problem: We say that an operator equation of the form (38) is *well posed* if (either \mathcal{A} maps \mathcal{H} onto \mathcal{H}' or $f \in \text{range}(\mathcal{A})$) and there exist positive constants c_A, C_A such that

$$c_A \|v\|_{\mathcal{H}} \leq \|\mathcal{A}v\|_{\mathcal{H}'} \leq C_A \|v\|_{\mathcal{H}} \quad \text{for all } v \in \mathcal{H} \quad (40)$$

Here \mathcal{H}' is the dual of \mathcal{H} endowed with the norm

$$\|w\|_{\mathcal{H}'} := \sup_{v \in \mathcal{H}} \frac{(v, w)}{\|v\|_{\mathcal{H}}} \quad (41)$$

and (\cdot, \cdot) is a dual form on $\mathcal{H} \times \mathcal{H}'$ (which is induced as before by the standard inner product in some pivot L_2 space).

Clearly (40) means that for any data f in the dual \mathcal{H}' – the range of \mathcal{A} – there exists a unique solution u , which depends continuously on the data f . Thus well-posedness refers to continuity with respect to a specific topology given by the *energy space* \mathcal{H} . It is not hard to show that in the case of Poisson's problem (36), (40) is a consequence of H^1 -ellipticity

$$\begin{aligned} (\nabla v, \nabla v) &\geq c \|v\|_{H^1(\Omega)}^2 := \|v\|_{L_2(\Omega)}^2 + \|\nabla v\|_{L_2(\Omega)}^2, \\ |(\nabla v, \nabla w)| &\leq C \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} \end{aligned} \quad (42)$$

which in turn follows from Poincaré's inequality. While in this special case the right space $\mathcal{H}' = H_0^1(\Omega)$ is easily recognized, the identification of a suitable \mathcal{H} such that (40) holds is sometimes a nontrivial task, an issue that will be taken up again later.

An important observation is that, once the mapping property (40) has been established, the Riesz basis property (30) for the energy space allows one to transform the original problem into an equivalent one which is now *well-posed* in the Euclidean metric. This is of particular importance in parameter dependent cases such as the Hilbert space \mathcal{H}_ϵ considered in the previous section.

Theorem 1. Suppose that $\mathcal{A} : \mathcal{H} \rightarrow \mathcal{H}'$ satisfies (40) and that Ψ is a Riesz basis for \mathcal{H} , that is (30) holds. Let A denote the standard representation of \mathcal{A} with respect to Ψ . Then (38) is equivalent to $Au = f$, where $u = \sum_{\lambda \in J} u_\lambda \psi_\lambda$, $f = ((\psi_\lambda, f))_{\lambda \in J}$. Moreover, A is boundedly invertible on ℓ_2 , that is,

$$c_\Psi^2 c_A \|v\|_{\ell_2} \leq \|Av\|_{\ell_2} \leq c_\Psi^2 C_A \|v\|_{\ell_2}, \quad \text{for all } v \in \ell_2 \quad (43)$$

Proof. By (30), one has for any $v = \sum_\lambda v_\lambda \psi_\lambda$

$$\begin{aligned} \|v\|_{\ell_2} &\leq c_\Psi^{-1} \|v\|_{\mathcal{H}} \leq c_\Psi^{-1} c_A^{-1} \|\mathcal{A}v\|_{\mathcal{H}'} \\ &\leq c_\Psi^2 c_A^{-1} \|((\psi_\lambda, \mathcal{A}v))_{\lambda \in J}\|_{\ell_2} = c_\Psi^2 c_A^{-1} \|Av\|_{\ell_2} \end{aligned}$$

where we have used (34). The reverse estimate follows analogously \square

3 EVOLUTION PROBLEMS — COMPRESSION OF FLOW FIELDS

We shall now address the problem class (1) of evolution equations. In many relevant instances, the evolution of the quantity $u(x, t)$ expresses a *conservation law* in the sense that for any test volume V in the domain Ω of interest,

$$\frac{d}{dt} \int_V u \, dx + \int_{\partial V} f(u) \cdot n \, dx = 0 \quad (44)$$

where $f(u)$ is the *flux function* and n denotes the outward normal on the boundary ∂V . When the solution is sufficiently smooth, (44) leads to a first order system of partial differential equations of the form

$$\partial_t u + \text{div}_x f(u) = 0 \quad (45)$$

which is said to be *hyperbolic* when the Jacobian matrix $Df(x)$ has for all x real eigenvalues with a full basis of eigenvectors. Hyperbolic systems of conservation laws are

used to model phenomena as diverse as traffic, information, and fluid flows. Perhaps the most well-known example is the system of Euler equations which model compressible fluid flow in terms of balance equations for mass, momentum, and energy. Such systems have to be complemented by suitable initial/boundary conditions. For simplicity, we shall assume pure initial data $u_0(x) = u(x, 0)$ with compact support.

Numerical methods for solving (44) are typically based on evolving cell averages $\bar{u}_C(t) := (\int_C u(x, t) \, dx)/|C|$, where C runs over a partition \mathcal{P} of the domain Ω into disjoint cells. In fact, the balance relation (44) also reads

$$\begin{aligned} \bar{u}_C(t + \Delta t) &= \bar{u}_C(t) + \frac{\Delta t}{|C|} B_C(t), \\ B_C(t) &:= \frac{1}{\Delta t} \int_t^{t+\Delta t} \int_{\partial C} f(u) \cdot n \, dx \, dt \end{aligned} \quad (46)$$

A *finite volume scheme* will mimic this time evolution by replacing the exact flux balance $B_C(t)$ by a numerical approximation computed from the current approximation of the exact cell average. More precisely, given a time step Δt , the scheme computes approximate values $u_C^n \approx u_C(n\Delta t)$ according to

$$u_C^{n+1} = u_C^n + \frac{\Delta t}{|C|} \sum_{C' \neq C} F_{C,C'}^n \quad (47)$$

where $F_{C,C'}^n$ is the numerical flux across the common boundary of C and an adjacent cell C' for the time interval $[n\Delta t, (n+1)\Delta t]$. This numerical flux typically depends on the values u_C^n and $u_{C'}^n$, and possibly on other neighboring values. We shall always assume that the scheme is *conservative*, that is, $F_{C,C'}^n = -F_{C',C}^n$. The initialization of the scheme uses the exact (or approximately computed) averages $u_C^0 := (\int_C u_0(x) \, dx)/|C|$ of the initial data u_0 .

Denoting the array of cell averages by $\mathbf{u}^n = (u_C^n)_{C \in \mathcal{P}}$, the finite volume scheme is thus summarized by a one step relation

$$\mathbf{u}^{n+1} = \mathbf{E} \mathbf{u}^n \quad (48)$$

where \mathbf{E} is a nonlinear discrete evolution operator. The computationally expensive part of this numerical method is the evaluation of the numerical fluxes $F_{C,C'}^n$, which is typically based on the (approximate) solution of local Riemann problems. An a priori error analysis of these classical numerical methods is only available to a limited extent. It refers to scalar problems, not to systems, and is rigorously founded only for uniform meshes. The proven error estimates are of low approximation orders like $h^{1/2}$ where h is the mesh size, that is, the maximal diameter of the cells.

It is well-known that the solutions of hyperbolic conservation laws exhibit a highly nonhomogeneous structure. Discontinuities in the solution can develop after finite time even for arbitrarily smooth initial data. So the solution exhibits regions of high regularity separated by regions of discontinuities (shocks). Capturing the singular effects in the solution by using classical discretizations based on uniform (or even quasi-uniform) partitions into cells would require a very fine resolution near the singularities and thus lead to enormous problem sizes. We see that the nature of the solution begs for the use of adaptive methods which would give finer resolution in the regions of shock discontinuities and maintain coarser resolution otherwise. The usual numerical approach is to generate such partitions adaptively. The difficulties in such an approach are to determine these regions and properly perform the time evolution on these inhomogeneous discretizations.

The analytic structure of solutions to (45) also points to possible advantages in using multiscale decompositions of the solution u in a numerical procedure. Because of the cancellation property (26), the coefficients of u would be small in those regions where the solution is smooth and would have significant size only near shocks. Thus, a multiscale decomposition would be excellent at identifying the regions of discontinuities by examining the size of the coefficients, and providing economical representations of the approximate solution at time $n\Delta t$ by an adapted set of wavelet coefficients $(d_{\lambda}^n)_{\lambda \in \Lambda^n}$. The approximate solution would therefore be given by

$$u_{\Lambda^n} = \sum_{\lambda \in \Lambda^n} d_{\lambda}^n \psi_{\lambda} \quad (49)$$

where the set Λ^n is allowed to vary with n . The main difficulty in this approach is the method of performing the evolution step strictly in terms of the wavelet coefficients. In other words, given Λ_n and the coefficients $(d_{\lambda}^n)_{\lambda \in \Lambda^n}$, how would we evolve on these data to obtain a good approximation at the next time step? This has led to the introduction of *dynamically adaptive schemes* in Maday, Perrier and Ravel (1991) in which the derivation of $(\Lambda^{n+1}, u_{\Lambda^{n+1}})$ from $(\Lambda^n, u_{\Lambda^n})$ typically goes in three basic steps:

- (i) **Refinement:** a larger set $\tilde{\Lambda}^{n+1}$ with $\Lambda^n \subset \tilde{\Lambda}^{n+1}$ is derived from a posteriori analysis of the computed coefficients $d_{\lambda}^n, \lambda \in \Lambda^n$.
- (ii) **Evolution:** a first numerical solution $u_{\tilde{\Lambda}^{n+1}} = \sum_{\lambda \in \tilde{\Lambda}^{n+1}} d_{\lambda}^{n+1} \psi_{\lambda}$ is computed from u_n and the data of the problem.
- (iii) **Coarsening:** the smallest coefficients of $u_{\tilde{\Lambda}^{n+1}}$ are thresholded, resulting in the numerical solution $u_{\Lambda^{n+1}} = \sum_{\lambda \in \Lambda^{n+1}} d_{\lambda}^{n+1} \psi_{\lambda}$ supported on the smaller set $\Lambda^{n+1} \subset \tilde{\Lambda}^{n+1}$.

A few words are in order concerning the initialization of the scheme: ideally, we can obtain an adaptive expansion u_{Λ^0} of the initial value data u_0 into a linear combination of wavelets by a thresholding procedure on its global expansion, that is,

$$u_{\Lambda^0} = \sum_{\lambda \in \Lambda^0} d_{\lambda}^0 \psi_{\lambda}, \quad \Lambda^0 := \{\lambda \text{ s.t. } \|d_{\lambda}^0 \psi_{\lambda}\|_{\mathcal{X}} \geq \eta\} \quad (50)$$

where \mathcal{X} is some prescribed norm in which we target to measure the error, η a prescribed threshold and $d_{\lambda}^0 := (u_0, \psi_{\lambda})$ are the wavelet coefficients of u_0 . In practice, we cannot compute all the values of these coefficients, and one thus needs a more reasonable access to a compressed representation. This is typically done through some *a priori* analysis of the initial value u_0 . In particular, if u_0 is provided by an analytic expression, or if we have some information on the local size of its derivatives, estimates on the decay of wavelet coefficients, such as (26), can be used to avoid the computation of most details which are below threshold. With such a strategy, we expect to obtain Λ^0 and $(d_{\lambda}^0)_{\lambda \in \Lambda^0}$ with a memory and computational cost which is proportional to $\#(\Lambda^0)$.

Then, assuming that at time $n\Delta t$, the approximate solution u_{Λ^n} has the form (49) for some set Λ^n of coefficients, the problem is thus both to select a correct set of indices Λ^{n+1} and to compute the new coefficients d_{λ}^{n+1} for $\lambda \in \Lambda^{n+1}$. As we have already explained, this is done by (i) refining Λ^n into an intermediate set $\tilde{\Lambda}^{n+1}$ which is well fitted to describing the solution at time $(n+1)\Delta t$, (ii) computing $u_{\tilde{\Lambda}^{n+1}}$ supported by $\tilde{\Lambda}^{n+1}$ and (iii) deriving $(u_{\Lambda^{n+1}}, \Lambda^{n+1})$ from $u_{\tilde{\Lambda}^{n+1}}$ by a thresholding process. The selection of the intermediate set $\tilde{\Lambda}^{n+1}$ should thus take into account the effect of the evolution operator \mathcal{E} on the sparse expansion (49), integrated between $n\Delta t$ and $(n+1)\Delta t$. Once a procedure for the refinement of Λ^n into $\tilde{\Lambda}^{n+1}$ has been prescribed, several strategies are available for computing $u_{\tilde{\Lambda}^{n+1}}$ from u_{Λ^n} , such as Petrov-Galerkin methods in Maday, Perrier and Ravel (1991) or collocation methods in Bertoluzza (1997). All these strategies are based on the computation of the inner products $(\mathcal{E}(u_{\Lambda^n}), \psi_{\lambda})$ for $\lambda \in \tilde{\Lambda}^{n+1}$ up to some precision. In the case where the evolution operator \mathcal{E} is linear, this amounts to a matrix-vector product, and one can make use of the sparse multiplication algorithm which will be discussed in Section 6. However, in many cases of interest, the evolution operator \mathcal{E} is nonlinear, making this computation more difficult and costly. Generally speaking, the discretization of nonlinear operators is a less simple task in the wavelet coefficient domain than in the physical domain.

In the following, we shall present a systematic approach which allows us to solve this problem, by a suitable combination of the representations of the numerical solution

by its wavelet coefficients and its physical values such as cell averages. This approach was first advocated by Ami Harten (Harten, 1993, 1995). The idea of Harten is to use multiscale decompositions where they do well – namely, in finding the discontinuities in the solution at a given time, and to use classical finite volume solvers, based on cell averages for the evolution step according to (47) and (51), since the properties of these solvers are well understood. To accomplish this, we need to build cell averages into our multiscale structure. This is easily accomplished (as is detailed below) by using multiscale bases that use characteristic functions of cells as the dual scaling functions. This means that at any given time step, one can view the numerical solution through one of two microscopes. The one is the decomposition as a sum of characteristic functions of cells (on the finest level of decomposition); the other is the multiscale decomposition. The first is good for the evolution; the second is good for identifying shocks and regions of smoothness. As described in Section 2.3, there are fast methods for transforming between the two sets of coefficients (scaling coefficients and multiscale coefficients).

Let us first amplify on our claim that cell averages lend themselves naturally to multiresolution techniques based on multilevel bases as described in Section 2.3. In fact, given a hierarchy of nested meshes and corresponding partitions $(\mathcal{P}_j)_{j \geq 0}$ of the flow domain, the cell averages u_C correspond directly to inner products $(u, \chi_C)/|C|$ which suggests that the L_1 -normalized functions $\chi_C/|C|$ for $C \in \mathcal{P}_j$ play the role of the dual scaling functions $\phi_{j,k}$. In complete analogy to (4), the indicator functions $\chi_C/|C|$ satisfy two-scale relations. The prediction operators have then the form (20) based on the refinement matrices $M_{j,0}$. Similarly to (4) one can construct Haar-like orthogonal bases ψ_{λ} . Here, as we have described in Section 2.3, one has a choice in the construction of the complement bases. We shall see that there is an advantage in having more vanishing moments in the dual basis than is provided by the classical Haar decomposition. Since Haar type bases have only first order vanishing moments, recall (12), one cannot expect a significant data compression. Therefore, one can use (22), (23) (with $\mathbf{K} = \mathbf{I}$) to raise the order of vanishing moments of the dual wavelets $\tilde{\psi}_{\lambda}$ as explained in (24) at the end of Section 2.3 (see Dahmen, Gottschlich-Müller and Müller, 2001; Müller, 2003). This amounts to changing the primal multiresolution spaces and in turn the primal wavelets ψ_{λ} while the dual scaling functions remain defined as $\chi_C/|C|$.

Given any array $\mathbf{v}_j = (v_C)_{C \in \mathcal{P}_j}$ of cell averages on the partition \mathcal{P}_j , we can transform this vector into the multiscale format $\mathbf{d}^j := (v_0, d_0, \dots, d_{J-1})$, where the arrays d_j encode detailed information needed to update the coarse cell averages in \mathbf{v}_j to \mathbf{v}_{j+1} on the next level of resolution. Thus the \mathbf{d} vectors correspond to the multiscale coefficients.

It is important to note that in generating the multiscale coefficients, we do not need explicit information on the multiscale basis functions. The transformation T_j that maps a cell average vector \mathbf{v}_j on to its multiscale decomposition \mathbf{d}^j , can be executed in an entirely discrete way as in the cascade algorithm of Section 2.3 (see also Arandiga, Donar and Harten, 1998, 1999; Carricer, Dahmen and Peña, 1996; Sweidens, 1996, 1998). To go the other way, from multiscale coefficients to the scaling coefficients, we again use the cascade structure. Recall that scaling coefficients \mathbf{v}_{j+1} , for a resolution level $j+1$, are obtained from the scaling coefficients \mathbf{v}_j at the coarser level in a two step process. The first is to predict \mathbf{v}_{j+1} by some rule (the lifting rule) and the second is to correct for the deviation in the prediction from the actual values. The deviation of the true coefficients \mathbf{v}_{j+1} from the predicted coefficients is given by the detail \mathbf{d}_{j+1} .

Our adaptive numerical scheme will be designed as a combination of a *reference* finite volume scheme which operates at the finest resolution level J according to

$$\mathbf{u}_j^{n+1} = \mathbf{E}_j \mathbf{u}_j^n \quad (51)$$

and of the transformations T_j and T_j^{-1} that relate the cell-average vector \mathbf{u}_j^n and its multiscale coefficients $(d_{\lambda}^n)_{\lambda \in \Lambda^{n,j-1}}$. In an adaptive context, we want to encode only a small relevant portion of this vector corresponding to the adaptive set Λ^n . Ideally, this set would correspond to the indices λ such that

$$\|d_{\lambda}^n \psi_{\lambda}\|_{\mathcal{X}} > \eta \quad (52)$$

where $\|\cdot\|_{\mathcal{X}}$ is the norm in which we plan to measure the error and η is some prescribed threshold. In practice, we precisely want to avoid the encoding of \mathbf{u}_j^n and of the full multiscale vector, and therefore we cannot invoke a thresholding procedure applied to the reference numerical solution. Therefore, we shall develop an adaptive strategy that iteratively computes some η -significant sets Λ^n and multiscale coefficients $(d_{\lambda}^n)_{\lambda \in \Lambda^n}$ which might differ from those obtained by thresholding \mathbf{u}_j^n . One of our goals is to keep track of the error between \mathbf{u}_j^n and the adaptive solution \mathbf{v}_j^n which is defined as the reconstruction on the finest partition \mathcal{P}_J from the details $(d_{\lambda}^n)_{\lambda \in \Lambda^n}$.

At this stage, a key observation is that a restricted multiscale vector $(d_{\lambda}^n)_{\lambda \in \Lambda^n}$ exactly encodes the cell averages on an *adaptive partition* $\mathcal{P}(\Lambda)$ which includes cells of different resolution levels $j = 0, \dots, J$ as illustrated in Figure 4, provided that the set Λ has a *graded tree structure*. Such a tree structure ensures that all the detail coefficients which are necessary to reconstruct the exact average on a cell $C \in \mathcal{P}(\Lambda)$ are contained in Λ . Note that a graded tree structure is not guaranteed on an adaptive set Λ produced

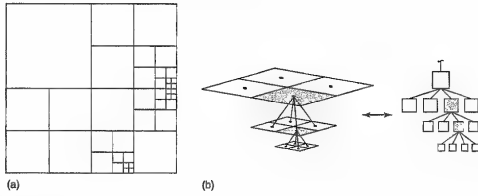


Figure 4. (a) Adaptive mesh (b) Tree.

by a thresholding procedure, yet it can be ensured by a suitable enlargement of Λ . In addition, it can be seen that the graded tree structure induces a grading property on the partition $\mathcal{P}(\Lambda)$ which essentially means that two adjacent cells differ at most by one resolution level. The concepts of tree structure and grading will also serve in Section 6 in the context of nonlinear variational problems. Another key observation is that the cost of the transformation T_Λ which maps the cell averages $(u_\lambda)_{\lambda \in \mathcal{P}(\Lambda)}$ onto the restricted multiscale vector $(d_\lambda)_{\lambda \in \Lambda}$ is of the order $\#\Lambda$ and similarly for the inverse transformation T_Λ^{-1} . A detailed description of such techniques and the design of appropriate data structures can be found in Müller (2003).

Based on this observation, we can propose the following adaptive scheme which follows the same principles as the dynamically adaptive scheme introduced in Maday, Perrier and Ravel (1991):

(1) *Initial values:* Apply the multiscale transform T_J to the initial cell averages u_J^0 to obtain the array of detail or wavelet coefficients $(d_\lambda^0)_{\lambda \in \Lambda^0}$ (including the cell averages on the coarsest level) for the time level $n=0$. Choose a threshold parameter $\eta > 0$ and set Λ^0 to be the smallest graded tree containing those λ such that $\|d_\lambda^0 \psi_\lambda\|_{L_1} > \eta$.

(2) *Predicting the significant indices on the next time level:* Given the η -significant tree Λ^n for the time level n and the details $(d_\lambda^n)_{\lambda \in \Lambda^n}$, predict a set $\tilde{\Lambda}^{n+1}$ that should contain the η -significant graded tree for time level $n+1$. We extend the detail vector by setting $d_\lambda^n = 0$ for $\lambda \in \tilde{\Lambda}^{n+1} \setminus \Lambda^n$ and we derive the cell averages $(v_\lambda^{n+1})_{\lambda \in \mathcal{P}(\tilde{\Lambda}^{n+1})}$ by applying the adaptive multiscale transform $T_{\tilde{\Lambda}^{n+1}}^{-1}$.

(3) *Time evolution step:* Compute the evolved cell averages $(v_\lambda^{n+1})_{\lambda \in \mathcal{P}(\tilde{\Lambda}^{n+1})}$ at time $n+1$, by some discrete evolution operator to be specified later. Of course it is important that this evolution can be done at less cost than would be necessary to evolve the uncompressed data.

(4) *Reverse transform and thresholding:* Apply the localized transform $T_{\tilde{\Lambda}^{n+1}}$ to $(v_\lambda^{n+1})_{\lambda \in \mathcal{P}(\tilde{\Lambda}^{n+1})}$ which yields an array of detail coefficients $(d_\lambda^{n+1})_{\lambda \in \tilde{\Lambda}^{n+1}}$. Set Λ^{n+1} to be the smallest graded tree containing those λ such that $\|d_\lambda^{n+1} \psi_\lambda\|_{L_1} > \eta$. Set $n+1$ to n and go to (2).

Any concrete realization of this scheme has to address the following issues.

- (1) Choice of the norm $\|\cdot\|_{L_1}$ and of the threshold parameter η ;
- (2) Strategy for predicting the set $\tilde{\Lambda}^{n+1}$;
- (3) Specification of the evolution operator.

Regarding (1), since the norm $\|\cdot\|_{L_1}$ will typically measure the deviation between the reference and adaptive solution u_J^0 and v_J^0 , a relevant choice for this norm should be such that we already have at our disposal an error estimate between the reference solution u_J^0 and the exact solution at time $n\Delta t$ in the same norm. As we shall see further, it will also be important that the reference scheme is stable with respect to such a norm. In the context of conservation laws, this limits us to the choice $X = L_1$. For a discrete vector u_J indexed by the finest partition \mathcal{P}_J , we define $\|u_J\|_{L_1}$ as the L_1 norm of the corresponding piecewise constant function on \mathcal{P}_J , that is,

$$\|u_J\|_{L_1} := \sum_{C \in \mathcal{P}_J} |C| |u_C| \quad (53)$$

Assuming that the ψ_λ are normalized in L_1 , it follows from the triangle inequality that the error e_n produced by discarding those multiscale coefficients of u^0 satisfying $\|d_\lambda \psi_\lambda\|_{L_1} \leq \eta$ is bounded by

$$e_n \leq \sum_{\{d_\lambda \psi_\lambda\|_{L_1} \leq \eta\}} \eta = \eta \#\{\lambda : \|d_\lambda \psi_\lambda\|_{L_1} \leq \eta\} \quad (54)$$

Since the above sum is limited to $|\lambda| \leq J-1$, we can derive the estimate

$$e_n \leq \#(\mathcal{P}_J) \eta \lesssim 2^{dJ} \eta \quad (55)$$

where d is the spatial dimension of the problem. It follows that a prescribed thresholding error δ can be obtained by using a threshold of the order

$$\eta \sim 2^{-dJ} \delta \quad (56)$$

Since the dual scaling functions and wavelets are normalized in L_1 , the primal scaling functions and wavelets are normalized in L_∞ so that $\|\psi_\lambda\|_{L_1} \sim 2^{-d|\lambda|}$. Therefore, the above strategy corresponds to applying to the coefficients d_λ a level dependent threshold $\eta_{|\lambda|}$ with

$$\eta_{|\lambda|} \sim 2^{d(J-|\lambda|)} \delta \quad (57)$$

Note however that the estimate (55) is somehow pessimistic since some thresholded coefficients d_λ could actually be much smaller than $\eta_{|\lambda|}$.

Concerning (2), an ideal prediction should take into account the action of the reference scheme E_J on the adaptive solution in the sense that the detail coefficients of $E_J v_J^0$ which are not contained in $\tilde{\Lambda}^{n+1}$ are guaranteed to be below the threshold. A strategy for constructing $\tilde{\Lambda}^{n+1}$ was proposed by Harten, based on a heuristic argument concerning the finite propagation speed of information in hyperbolic problems. Basically, $\tilde{\Lambda}^{n+1}$ is formed as the union of certain fixed neighborhoods (on the same or at most one higher scale) of the elements in Λ^n . Recently, at least for scalar problems, a rigorous analysis has been presented in Cohen *et al.* (2002) which gives rise to sets $\tilde{\Lambda}^{n+1}$ that are *guaranteed* to fulfill the above prescription. In this case, the neighborhoods are allowed to depend in a more precise way on the size of the elements in Λ^n . In practical experiments, Harten's simpler choice seems to have worked so far well enough though.

Turning to (3), several strategies are available for evolving the cell averages $(v_\lambda^{n+1})_{\lambda \in \mathcal{P}(\tilde{\Lambda}^{n+1})}$ into $(v_\lambda^{n+2})_{\lambda \in \mathcal{P}(\tilde{\Lambda}^{n+2})}$. The first one consists in computing the effect on these averages of the exact application of the reference scheme E_J to the adaptive solution v_J^0 reconstructed on the fine grid. A key observation is that since we are only interested in the averages of $E_J v_J^0$ on the adaptive partition $\mathcal{P}(\tilde{\Lambda}^{n+1})$, the numerical fluxes which need to be computed are only those between the adjacent fine cells such that their interface lies on the edge of a cell of the adaptive partition. In the original concept proposed by Harten, this idea was exploited in order to obtain CPU savings on the number of flux evaluation, with the solution encoded in its nonadaptive form v_J^0 . In Cohen *et al.* (2002), it was shown that the computation

of the needed fluxes can be performed from the adaptive data $(v_\lambda^{n+1})_{\lambda \in \mathcal{P}(\tilde{\Lambda}^{n+1})}$ without the need of performing the reconstruction of the entire v_J^0 . This information can indeed be acquired by local reconstruction. However, in several space dimensions, the resulting computational complexity, although still lower than that for the fully refined partitions, is suboptimal. A second, more economical strategy is to employ the finite volume stencil of the uniform partition but for the currently local level of resolution. This makes use of the local quasi-uniformity of the mesh which can be made locally uniform by subdividing neighboring cells of lower generation. The gradedness of the partitions ensures that the subdivisions need only have depth one. In numerical experiments, this strategy turns out to work well when using higher order finite volume schemes in connection with corresponding higher order multiscale decompositions, here corresponding to the higher order vanishing moments (see e.g. Müller, 2003).

One of the good features of the adaptive approach that we have described is the possibility to monitor the error between the reference and adaptive numerical solution by a proper tuning of the threshold parameter. Here, we consider the evolution strategy that amounts to computing exactly the averages of $E_J v_J^0$ on the adaptive partition $\mathcal{P}(\tilde{\Lambda}^{n+1})$. It follows that we can write

$$\|u_J^{n+1} - v_J^{n+1}\|_{L_1} = \|E_J u_J^n - E_J v_J^n\|_{L_1} + p_n + t_n \quad (58)$$

where

$$p_n := \sum_{\lambda \notin \tilde{\Lambda}^{n+1}} \|d_\lambda(E_J v_J^n) \psi_\lambda\|_{L_1} \quad (59)$$

and

$$t_n := \sum_{\lambda \in \tilde{\Lambda}^{n+1} \setminus \Lambda^{n+1}} \|d_\lambda(E_J v_J^n) \psi_\lambda\|_{L_1} \quad (60)$$

respectively denote the errors resulting from the restriction to the predicted set $\tilde{\Lambda}^{n+1}$ and to the set Λ^{n+1} obtained by thresholding. According to our previous remarks, these errors can be controlled by some prescribed δ provided that we use the level dependent threshold $\eta_{|\lambda|} \sim 2^{d(J-|\lambda|)} \delta$. Assuming in addition that the reference scheme is L_1 -stable in the sense that for all u_J and v_J ,

$$\|E_J u_J - E_J v_J\|_{L_1} \leq (1 + C\Delta t) \|u_J - v_J\|_{L_1} \quad (61)$$

we thus obtain

$$\|u_J^{n+1} - v_J^{n+1}\|_{L_1} \leq \|E_J u_J^n - E_J v_J^n\|_{L_1} + 2\delta \quad (62)$$

which yields the estimate at time $T = n\Delta t$

$$\|u_J^n - v_J^n\|_{L_1} \leq C(T)\delta \quad (63)$$

Therefore, if the reference numerical scheme is known to provide accuracy $\epsilon = \epsilon_J$ on level J , it is natural to choose δ such that $n\delta \sim \epsilon$. In many practical instances, however, this estimate turns out to be too pessimistic in the sense that thresholding and refinement errors do not really accumulate with time, so that δ and the threshold η can be chosen larger than the value prescribed by this crude analysis. A sharper analysis of the error between the adaptive and reference solution is still not available.

An adaptive solver based on the above concepts has been developed and implemented by S. Müller. It also incorporates implicit time discretizations. A detailed account can be found in Müller (2003). The recently developed new flow solver QUADFLOW for hyperbolic conservation laws and for the Navier Stokes equations for compressible flows is based on these adaptive multiresolution techniques, on a finite volume discretization that can cope with hanging nodes and on a mesh generator based on block partitions. Each block corresponds to a B-spline based parametric mapping that allows a flexible mesh refinement through point evaluations of the B-spline representation. An outline of the scheme and extensive numerical tests can be found in Bramkamp *et al.* (2001) and Ballmann, Bramkamp and Müller (2000). The numerical examples provided by S. Müller and F. Bramkamp should give an impression of the performance of such techniques. The first example in Figure 5 shows the results for an Euler computation concerning a flow at Mach 0.95 at an angle of attack $\alpha = 0$ around a bench mark NACA0012 profile. Here the main objective is to test the resolution of shock interactions even at a large distance from the airfoil. The mesh has approximately 5×10^4 cells as opposed to an estimated number of 7×10^7 cells needed by a uniformly refined mesh for a comparable target accuracy. A close up of Figure 5(b) is displayed in Figure 6.

Figure 7 shows a series of adaptive refinements again for an Euler computation for a flow around a BAC 3-11/RES/30/21-Profile at $M = 0.85$ and an angle of attack $\alpha = 0^\circ$. This test illustrates the reliable detection even of small shocks here in the lower region of the nose. Further detailed numerical studies for stationary problems such as moving wings, shock-bubble interactions, analogous studies for viscous flows and boundary layer resolution can be found in Bramkamp *et al.* (2001), Ballmann, Bramkamp and Müller (2000), and Müller (2003).

3.1 Concluding remarks

The above framework is an example where multiscale bases for realistic geometries are conveniently realized. In spite of the promising numerical results, it should be stressed

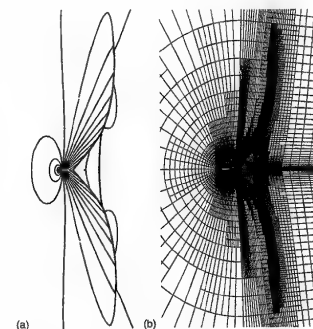


Figure 5. (a) Pressure distribution (b) Adaptive mesh.

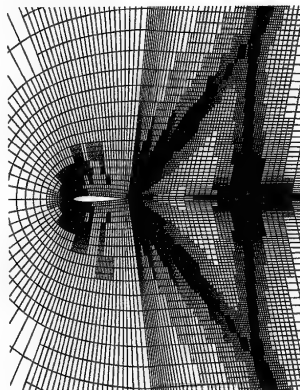


Figure 6. Adaptive mesh – close up.

though that many principal questions remain open, due to a number of obstructions. First, the current understanding of error analysis for hyperbolic problems is much less

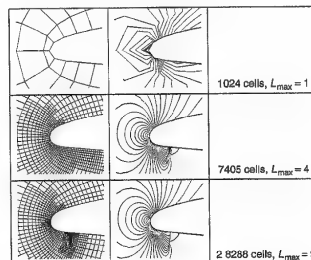


Figure 7. Adaptive mesh refinement.

developed than for elliptic problems, partly due to the nature of the relevant function spaces. On one hand, there are only poor a priori estimates that could serve as a benchmark. The central point of view is a *perturbation analysis*. The overall attainable accuracy is fixed by the a priori choice of a highest level J of spatial resolution. All subsequent attempts aim at preserving the accuracy offered by a uniformly refined discretization with that resolution at possibly low cost. Thus whatever information is missed by the reference scheme cannot be recovered by the above adaptive solver.

On the other hand, the multiscale techniques did not unfold their full potential; one makes use of the cancellation properties of wavelet bases but not of the norm equivalences between wavelet coefficients and functions. Thus the primary focus here is on the compression of the conserved variables, that is, on the *sparse approximation* of functions based on cancellation properties. This does so far not provide any estimates that relate the achieved accuracy ϵ to the size of the η -significant trees. This question will be addressed later, again in a different context, where stronger basis properties allow one to exploit not only the sparse approximation of functions but also that of the involved operators.

4 BOUNDARY INTEGRAL EQUATIONS — MATRIX COMPRESSION

A variety of problems in elasticity, fluid flow or electromagnetism lead to a formulation in terms of *boundary*

integral equations falling into the category (2). In principle, this is a feasible approach when the Green's function of the underlying (linear) partial differential equation is known explicitly. This is a particularly tempting option when the original formulation via a PDE refers to an exterior and hence unbounded domain since the corresponding boundary integral formulation lives on a compact manifold of lower spatial dimension. Wavelet concepts have had a significant impact on this problem area (see Dahmen, Harbrecht and Schneider, 2002; Dahmen, Pröddorf and Schneider, 1994; Harbrecht, 2001; Lage and Schwab, 1998; Lage, 1996; von Petersdorff and Schwab, 1996; von Petersdorff and Schwab, 1997; Schneider, 1998), primarily in finding sparse and efficient approximations of potential operators. We shall describe this in the following simple setting.

4.1 Classical boundary integral equations

Let Ω^- be again a bounded domain in \mathbb{R}^d ($d \in \{2, 3\}$) and consider Laplace's equation

$$-\Delta u = 0, \text{ on } \Omega, \quad (\Omega = \Omega^- \text{ or } \Omega^+ := \mathbb{R}^3 \setminus \Omega^-) \quad (64)$$

subject to the boundary conditions

$$w = f \quad \text{on } \Gamma := \partial\Omega^- \quad (w(x) \rightarrow 0, |x| \rightarrow \infty \text{ when } \Omega = \Omega^+) \quad (65)$$

Of course, the unbounded domain Ω^+ poses an additional difficulty in the case of such an *exterior* boundary value problem. A well-known strategy is to transform (64), (65) into a *boundary integral equation* that lives only on the manifold $\Gamma = \partial\Omega$. There are several ways to do that. They all involve the fundamental solution $\mathcal{E}(x, y) = 1/(4\pi|x - y|)$ of the Laplace operator which gives rise to the *single layer potential operator*

$$(\mathcal{A}u)(x) := (\mathcal{V}u)(x) := \int_{\Gamma} \mathcal{E}(x, y)u(y) d\Gamma_y, \quad x \in \Gamma \quad (66)$$

One can then show that the solution u of the first kind integral equation

$$\mathcal{V}u = f \quad \text{on } \Gamma \quad (67)$$

provides the solution w of (64) through the representation formula

$$w(x) = \int_{\Gamma} \mathcal{E}(x, y)u(y) d\Gamma_y, \quad x \in \Omega \quad (68)$$

An alternative way uses the double layer potential

$$\begin{aligned} (\mathcal{K}v)(x) &:= \int_{\Gamma} \frac{\partial}{\partial n_y} \mathcal{E}(x, y) v(y) d\Gamma_y \\ &= \int_{\Gamma} \frac{1}{4\pi} \frac{n_y^T(x-y)}{|x-y|^3} v(y) d\Gamma_y, \quad x \in \Gamma \end{aligned} \quad (69)$$

where n_y is the outward normal to Γ at $y \in \Gamma$. Now the solution of the second kind integral equation

$$\mathcal{A}u := (\tfrac{1}{2} \pm \mathcal{K})u = f \quad (\Omega = \Omega^\pm) \quad (70)$$

gives the solution to (64) through

$$w(x) = \int_{\Gamma} \mathcal{K}(x, y) u(y) d\Gamma_y \quad (71)$$

One way of reformulating (64) with Neuman boundary conditions $\partial w / \partial n = g$ on Γ , where $\int_{\Gamma} g(x) d\Gamma_x = 0$, is offered by the so-called hypersingular operator $(\mathcal{W}v)(x) := -(\partial/\partial n_x) \int_{\Gamma} (\partial/\partial n_y) \mathcal{E}(x, y) d\Gamma_y$. Now the solution of

$$\mathcal{A}u = \mathcal{W}u = g \quad \text{on } \Gamma \quad (72)$$

leads to the solution of the Neuman problem for Laplace's equation through the representation formula (71), where the constraint $\int_{\Gamma} u(x) d\Gamma_x = 0$ is imposed in the case of an interior problem to ensure uniqueness.

4.2 Quasi-sparsity of wavelet representations

We have encountered so far two classes of operators. The first case such as (36) concerns differential operators that are local in the sense that $\langle \psi_\lambda, \mathcal{A}\psi_\nu \rangle = 0$ whenever $S_\lambda \cap S_\nu = \emptyset$. Note that the wavelet representation A is not sparse in the classical sense since basis functions from different levels may overlap. In fact, it is easy to see that the number of entries in a principal section of A of size N contains $O(N \log N)$ entries. However, we shall see that many of these entries are so small in modulus that they can be neglected in a matrix-vector multiplication without perturbing the result too much. The point of focus in this section is that this even holds true for the second class of global operators, which are roughly speaking inverses of differential operators such as the above boundary integral operators. They all share the property that they (or at least their global part) are of the form

$$(\mathcal{A}u)(x) = \int_{\Gamma} K(x, y) u(y) d\Gamma_y \quad (73)$$

where for a given domain or manifold Γ the kernel $K(x, y)$ is smooth except on the diagonal $x = y$ and satisfies the decay conditions.

$$|\partial_x^\alpha \partial_y^\beta K(x, y)| \lesssim \text{dist}(x, y)^{-(d+2\alpha+|\beta|)} \quad (74)$$

By (39), the entries of A are in this case given by

$$\begin{aligned} A_{\lambda, \nu} &= \langle K, \psi_\lambda \otimes \psi_\nu \rangle_{\Gamma \times \Gamma} = \int_{\Gamma} \int_{\Gamma} K(x, y) \\ &\quad \times \psi_\lambda(x) \psi_\nu(y) d\Gamma_x d\Gamma_y \end{aligned} \quad (75)$$

Although none of the entries $A_{\lambda, \nu}$ will generally be zero, many of them are very small in modulus as specified by the following classical estimate (see e.g. Dahmen, Pröbldorf and Schneider, 1994; von Petersdorff and Schwab, 1996, 1997).

Theorem 2. Suppose that the kernel K is of the above form and that $\mathbf{D}^{-d}\Psi$ is a Riesz-basis for H^s for $-\tilde{\gamma} < s < \gamma$ (see (32)) and has cancellation properties (see (26)) of order \tilde{m} . Moreover, assume that A given by (73) has order $2t$ and satisfies for some $r > 0$

$$\|A v\|_{H^{-t+r}} \lesssim \|v\|_{H^{-t+r}}, \quad v \in H^{t+r}, 0 \leq |a| \leq r \quad (76)$$

Then, for any $\sigma > 0$ such that $0 < \sigma \leq \min\{r, d/2 + \tilde{m} + t\}$, $t + \sigma < \gamma$, and $t - \sigma > -\tilde{\gamma}$, one has

$$2^{-(|t|+|\lambda|)t} |\langle \psi_\lambda, \mathcal{A}\psi_\nu \rangle| \lesssim \frac{2^{-||\lambda|-|\nu||\sigma}}{(1+2^{\min(|\lambda|, |\nu|)} \text{dist}(S_\lambda, S_\nu))^{d+2\tilde{m}+2t}} \quad (77)$$

Thus the entries of the wavelet representation of operators of the above type exhibit a polynomial spatial decay, depending on the order of cancellation properties, and an exponential scalewise decay, depending on the regularity of the wavelets.

For a proof of such estimates, one distinguishes two cases. When $\text{dist}(S_\lambda, S_\nu) \lesssim 2^{-\min(|\lambda|, |\nu|)}$ one can use the continuity properties (76) in combination with the norm equivalences (32) to show that

$$|\langle \psi_\lambda, \mathcal{A}\psi_\nu \rangle| \leq 2^{(|\lambda|+|\nu|)2\sigma} 2^{d|\nu|-|\lambda|D} \quad (78)$$

(see Dahlke et al. (1997) for more details).

On the other hand, when $\text{dist}(S_\lambda, S_\nu) \gtrsim 2^{-\min(|\lambda|, |\nu|)}$, the wavelets are integrated against smooth parts of the kernel K . One can then exploit the cancellation properties for both wavelets to obtain the bound

$$|\langle \psi_\lambda, \mathcal{A}\psi_\nu \rangle| \lesssim \frac{2^{-(|\lambda|-|\nu|)(d/2+\tilde{m})}}{(\text{dist}(S_\lambda, S_\nu))^{d+2\tilde{m}+2t}} \quad (79)$$

(see e.g. Dahmen, Pröbldorf and Schneider (1994), Dahmen and Stevenson (1999), von Petersdorff and Schwab (1996), and von Petersdorff and Schwab (1997) for more details). The decay estimate (77) follows then from (78) and (79).

However, the above argument for the case of overlapping supports is rather crude. Instead one can use the so-called second compression due to Schneider, (Schneider, 1998). In fact, when $|\lambda| \gg |\nu|$ and when S_λ does not intersect the singular support of ψ_ν , then $\mathcal{A}\psi_\nu$ is smooth on the support of ψ_λ , and one can again use the cancellation property of ψ_λ . Denoting by S_ν^* the singular support of (the lower level wavelet) ψ_ν , this leads to

$$|\langle \psi_\lambda, \mathcal{A}\psi_\nu \rangle| \lesssim \frac{2^{1/2} 2^{-|\nu|(\tilde{m}-d/2)}}{\text{dist}(S_\lambda, S_\nu^*)^{d+2\tilde{m}}} \quad (80)$$

Estimates of the type (77), (78), and (80) provide the basis of matrix compression strategies that aim at replacing the wavelet representation of an operator by a sparsified perturbation that can be used to expedite the numerical solution of corresponding linear systems.

4.3 Weak formulations and Galerkin schemes

As in the case of Poisson's equation (36) we are dealing again with an operator equation

$$\mathcal{A}u = f \quad (81)$$

this time of the type (66), (70), or (72). A classical approach to solving such an equation numerically is to return again to a proper weak formulation on which to base a Galerkin discretization. As before, the key is to identify first a suitable (Hilbert) space \mathcal{H} such that the variational formulation

$$a(v, u) := (v, \mathcal{A}u) = (v, f) \quad \text{for all } v \in \mathcal{H} \quad (82)$$

is well-posed in the sense of (40). In terms of the operator \mathcal{A} , this can be rephrased by saying that \mathcal{A} is boundedly invertible as a mapping from \mathcal{H} onto \mathcal{H}' , which will often be referred to as mapping property.

All the above examples can be shown to fall into this framework (see e.g. Kress, 1989). The single layer potential is symmetric positive definite on the Sobolev space $\mathcal{H} := H^{-1/2}(\Gamma)$ whose dual is $\mathcal{H}' = H^{1/2}(\Gamma)$, that is,

$$a(v, v) = (v, v) \gtrsim \|v\|_{H^{-1/2}(\Gamma)}^2 \quad \text{for all } v \in H^{-1/2}(\Gamma) \quad (83)$$

which is easily seen to imply (40).

The double layer potential is known to be compact when Γ is a C^2 manifold in which case the kernel is weakly singular. In general, the appropriate energy space

is $\mathcal{H} = L_2(\Gamma) = \mathcal{H}'$. Despite the lack of symmetry, one can show that the bilinear form $a(\cdot, \cdot)$ is coercive and that (40) holds with $\mathcal{H} = L_2(\Gamma)$.

The hypersingular operator \mathcal{W} , in turn, is strongly singular with energy space $\mathcal{H} = H^{1/2}(\Gamma)$ (i.e. $\mathcal{H}' = H^{-1/2}(\Gamma)$), respectively $\mathcal{H} = H^{1/2}(\Gamma)/\mathbb{R}$ in the case of an interior problem. Again, since it is then symmetric positive definite and (40) follows.

According to the shifts caused by these operators in the Sobolev scale, $\mathcal{A} : H^s(\Gamma) \rightarrow H^{-s}(\Gamma)$, the single layer potential, double layer potential and hypersingular operator have order $2t = -1, 0, 1$, respectively.

The (conforming) Galerkin method (for any operator equation (64)) consists now in choosing a finite dimensional space $S \subset \mathcal{H}$ and determining $u_S \in S$ such that

$$a(v, u_S) = (v, f) \quad \text{for all } v \in S \quad (84)$$

Such a scheme is called (\mathcal{H}) -stable (for a family S of increasing spaces $S \in \mathcal{S}$) if (40) holds on the discrete level, uniformly in $S \in \mathcal{S}$. In other words, denoting by P_S any \mathcal{H} -bounded projector onto S , we need to ensure that (40) holds with A replaced by $P_S^* A P_S$, uniformly in $S \in \mathcal{S}$, where P_S^* is the adjoint of P_S . This is trivially the case for any subspace $S \subset \mathcal{H}$ when A is symmetric positive definite. In the coercive case, one can show that Galerkin discretizations are stable for families S of trial spaces that satisfy certain approximation and regularity properties formulated in terms of direct and inverse estimates, whenever the level of resolution is fine enough (see e.g. Dahmen, Pröbldorf and Schneider, 1994).

Once this homework has been done, what remains is choosing a basis for S by which (84) is turned into a linear system of equations. The unknowns are the coefficients of u_S with respect to the chosen basis.

The obvious advantage of the boundary integral approach is the reduction of the spatial dimension and that one has to discretize in all cases only on bounded domains. On the other hand, one faces several obstructions:

- (i) Whenever the order of the operator is different from zero (e.g. for $A = \mathcal{V}$), the problem of growing condition numbers arises because the operator treats high frequency components differently from slowly varying ones. In general, if an operator has order $2t$, the spectral condition numbers of the stiffness matrices grow like $h^{-2|t|}$, where h reflects the spatial resolution (e.g. the mesh size) of the underlying discretization.
- (ii) Discretizations lead in general to densely populated matrices. This severely limits the number of degrees of freedom when using direct solvers. But iterative techniques are also problematic, due to the fact that

the cost of each matrix/vector multiplication increases with the square of the problem size.

One possible strategy to overcome these obstructions will be outlined next.

4.4 Wavelet-Galerkin methods

We adhere to the above setting and consider the operator equation (81), where A has the form (73), (74) and satisfies (40) for $\mathcal{H} = H^1(\Gamma)$. Suppose now that we have a wavelet basis Ψ which is a Riesz basis for $H^1(\Gamma)$, constructed along the lines from Section 2.3, with the corresponding multiresolution sequence of spaces S_j spanned by all wavelets ψ_{λ} , $|\lambda| < j$, of level less than j . We point out next how the use of the spaces S_j as trial spaces in Galerkin discretizations can help to cope with the above obstructions. To this end, let $A_j := ((\psi_{\lambda}, A\psi_{\nu}))_{|\lambda|, |\nu| < j}$ denote the stiffness matrix of A with respect to the (finite) wavelet basis of the trial space S_j . Thus (84) takes the form

$$A_j u_j = f_j, \quad f_j := ((\psi_{\lambda}, f))_{|\lambda| < j} \quad (85)$$

The first observation concerns obstruction (i) above (see e.g. Dahmen and Kunoth, 1992; Dahmen, Pröbldorf and Schneider, 1994).

Remark 2. If the Galerkin discretizations are H^1 -stable for $\mathcal{S} = \{S_j\}_{j \in \mathbb{N}_0}$, then the spectral condition numbers of A_j are uniformly bounded.

In fact, when the bilinear form $a(\cdot, \cdot)$, defined in (82), is symmetric and H^1 -elliptic, so that A is symmetric positive definite, the spectrum of A_j is contained in the convex hull of the spectrum of A , so that the assertion follows immediately from Theorem 1. Since under this assumption Galerkin discretizations are always stable for any choice of subspaces, this is a special case of the above claim. In the general case, the argument is similar to that in the proof of Theorem 1. In fact, Galerkin stability means that $\|A_j^{-1}\|_{\ell_2 \rightarrow \ell_2} \lesssim 1$, which ensures the existence of a constant \tilde{c} such that $\tilde{c}\|v_j\|_{\ell_2} \leq \|A_j v_j\|_{\ell_2}$ for any $v_j \in S_j$ with coefficient vector v_j . Moreover, by (43),

$$\begin{aligned} \|A_j v_j\|_{\ell_2} &= \|(\psi_{\lambda}, A v_j)_{|\lambda| < j}\|_{\ell_2} \\ &\leq \|A v_j\|_{\ell_2} \leq C_0^2 C_{\mathcal{A}} \|v_j\|_{\ell_2} \end{aligned}$$

which confirms the claim.

This observation applies to all our above examples of boundary integral operators. In fact, \mathcal{V} and \mathcal{W} are elliptic and the coercivity in the case (70) of the double layer

potential ensures that (for $j \geq j_0$ large enough) the Galerkin discretizations are also stable in this case (see e.g. Dahmen, Pröbldorf and Schneider, 1994).

Thus, a proper choice of wavelet bases for the respective energy space deals with obstruction (i) not only for the second kind integral equations with zero order operators but also essentially for all classical potential operators. This is, for instance, important in the context of transmission problems.

Of course, the preconditioning can be exploited in the context of iterative solvers for (85) only if the cost of a matrix/vector multiplication can be significantly reduced below the square of the problem size. First, note that, due to the presence of discretization errors, it is not necessary to compute a matrix/vector multiplication *exactly* but it would suffice to *approximate* it within an accuracy tolerance that depends on the current discretization error provided by S_j . Thus, one faces the following central

Task: Replace as many entries of A_j as possible by zero so as to obtain a perturbed matrix \tilde{A}_j with the following properties for all the above operator types:

- (i) The \tilde{A}_j have still uniformly bounded condition numbers when the level j of resolution grows.
- (ii) The solutions \tilde{u}_j of the perturbed systems $\tilde{A}_j \tilde{u}_j = f_j$ have still the same order of accuracy as the solutions u_j of the unperturbed systems (85), *uniformly* in j .
- (iii) Find efficient ways of computing the nonzero entries of \tilde{A}_j .

These issues have been addressed in a number of investigations (see e.g. Dahmen, Pröbldorf and Schneider, 1994; Dahmen, Harbrecht and Schneider, 2002; von Petersdorff and Schwab, 1996, 1997; von Petersdorff, Schneider and Schwab, 1997; Harbrecht, 2001; Schneider, 1998). We shall briefly outline the current state of the art as reflected by Harbrecht (2001) and Dahmen, Harbrecht and Schneider (2002). The key is a suitable, level-dependent thresholding strategy based on the a priori estimates (79) and (80). It requires a sufficiently high order of cancellation properties, namely $\tilde{m} > m - 2t$ where m is the approximation order provided by the multiresolution spaces S_j . Thus, whenever A has nonpositive order (such as the single and double layer potential operator), one must have $\tilde{m} > m$, ruling out orthonormal wavelets (in L_2). Given that Ψ meets this requirement, and considering a fixed highest level J of resolution, fix parameters $a, a' > 1$ and $m' \in (m, \tilde{m} + 2t)$ and define the cut-off parameters (see Dahmen, Harbrecht and Schneider, 2002; Harbrecht, 2001; Schneider, 1998)

$$c_{l,j} := a \max \left\{ 2^{-\min(l,j)}, 2^{2J(m'-1)-(j+J)(m'+\tilde{m})/(2(\tilde{m}+t))} \right\}$$

$$c'_{l,j} := a' \max \left\{ 2^{-\max(l,j)}, 2^{2J(m'-1)-(j+J)m' - \max(j,J)\tilde{m}/(\tilde{m}+2t)} \right\} \quad (86)$$

Then the a priori compression of A_j is given by

$$(\tilde{A}_j)_{\lambda, \nu} := \begin{cases} 0, & \text{dist}(S_{\lambda}, S_{\nu}) > c_{|\lambda|, |\nu|} \text{ and } |\lambda|, |\nu| \geq j_0, \\ 0, & \text{dist}(S_{\lambda}, S_{\nu}) \lesssim 2^{-\min(|\lambda|, |\nu|)} \text{ and} \\ & \text{dist}(S'_{\lambda}, S'_{\nu}) > c'_{|\lambda|, |\nu|} \text{ if } |\nu| > |\lambda| \geq j_0 - 1, \\ & \text{dist}(S_{\lambda}, S'_{\nu}) > c'_{|\lambda|, |\nu|} \text{ if } |\lambda| > |\nu| \geq j_0 - 1, \\ (A_j)_{\lambda, \nu}, & \text{otherwise} \end{cases} \quad (87)$$

The first line is the classical 'first compression' based on (79) when the wavelets have disjoint supports with a distance at least the diameter of the larger support. The number of nonzero entries that remain after this compression is of the order $N_j \log N_j$ where $N_j := \dim S_j \sim 2^{(d-1)j}$ when $d-1$ is the dimension of Γ . The second line reflects the 'second compression' due to Schneider, which discards entries for wavelets with overlapping support (Schneider, 1998). More importantly, this affects also those entries involving the scaling functions on the coarsest level j_0 . It has a significant effect already when, due to a complicated geometry, the coarsest level already involves a relatively large number of basis functions. Asymptotically, it removes the log factor in the count of the nonzero entries of \tilde{A}_j .

A sophisticated perturbation analysis, whose main ingredients are the a priori estimates (79), (80) based on the cancellation properties of Ψ , the norm equivalences (30), and suitable versions of the Schur lemma, yields the following result (Dahmen, Harbrecht and Schneider, 2002; Harbrecht, 2001).

Theorem 3. The compressed matrices \tilde{A}_j , given by (87), have uniformly bounded condition numbers. The number of nonzero entries in \tilde{A}_j is of the order N_j , uniformly in J . Moreover, the solution \tilde{u}_j exhibits optimal discretization error estimates in the energy norm

$$\|u - \tilde{u}_j\|_{H^1(\Gamma)} \lesssim 2^{J(m-m')} \|u\|_{H^1(\Gamma)}, \quad J \rightarrow \infty \quad (88)$$

This result says that the above compression strategy is asymptotically optimal. In comparison with earlier versions, the removal of log-factors even offers a strictly linear complexity. Note that for operators of negative order, the relatively high computational efforts for Galerkin discretizations, due to the double integrals, pay off through the high order, $m + |t|$.

The remaining crucial question concerns the complexity of computing the compressed matrices \tilde{A}_j . A detailed

analysis of this issue can be found in Harbrecht (2001); see also Dahmen, Harbrecht and Schneider (2002). The main facts can be summarized as follows:

Of course, the entries of \tilde{A}_j cannot be computed exactly but one has to resort to quadrature. The following nice observation from Harbrecht (2001) tells us how much computational effort can be spent on the entry $(A_j)_{\lambda, \nu}$ so as to keep the overall complexity of computing an approximation to \tilde{A}_j proportional to the system size N_j .

Theorem 4. The complexity of approximately computing the nonzero entries of \tilde{A}_j is $O(N_j)$, provided that for some $\alpha > 0$ at most $O((J - (|\lambda| + |\nu|)/2)^\alpha)$ operations are spent on the computation of a nonzero coefficient $(\tilde{A}_j)_{\lambda, \nu}$.

Next, in analogy to the compression estimates, one can ask which further perturbation is allowed for the approximate calculation of the entries of \tilde{A}_j so as to retain the above optimal convergence rates (Harbrecht, 2001; Dahmen, Harbrecht and Schneider, 2002).

Theorem 5. If the quadrature error for $(\tilde{A}_j)_{\lambda, \nu}$ is bounded by

$$\delta \min \left\{ 2^{-(d-1)(|\lambda|+|\nu|)/2}, 2^{-(d-1)(J-(|\lambda|+|\nu|)/2)(m' - t)/(m+1)} \right\} \times 2^{2Jt} 2^{-2m'(J-(|\lambda|+|\nu|)/2)}$$

for some fixed $\delta < 1$, then the (perturbed) Galerkin scheme is stable and converges with optimal order (88).

The main result now is that the accuracy bounds in Theorem 5 can be met by a sophisticated adapted quadrature strategy whose computational complexity remains also in the operations budget given by Theorem 4. Thus, in summary, one obtains a fully discrete scheme that exhibits asymptotically optimal computational complexity that remains proportional to the problem size. This is illustrated below by some numerical examples.

Remark 3. Another approach to matrix compression, first pointed out in Beylkin, Coifman and Rokhlin (1991), uses the so-called *nonstandard form* of the operator $A_S := P_S^* A P_S$, which involves a telescoping expansion for A_S but is not a *representation* of A_S in the strict sense. It consists of blocks whose entries involve only basis functions (wavelets and scaling functions) of the same level, which may simplify their computation in comparison to the standard form. On the other hand, the nonstandard form does not support preconditioning when dealing with operators of order different from zero and is therefore restricted to problems of the type (70). Due to the presence of scaling function coefficients, it also does not allow us to combine matrix compression together with function

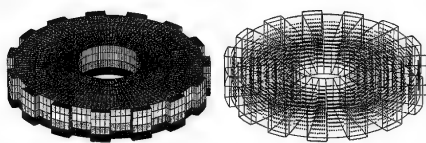


Figure 8. The surface mesh and the evaluation points x_i of the potential. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

compression. We shall point out later that this is indeed supported by the standard form.

4.5 Numerical tests

The following example has been provided by Harbrecht. An interior Dirichlet problem for the Laplacian is solved by the indirect approach. We use both, the Fredholm integral equation of the first kind based on the single layer operator and the Fredholm integral equation of the second kind based on the double layer potential operator. Both approaches yield a density u , from which one derives the solution in the domain via potential evaluation, that is, by applying the single layer operator and double layer operator, respectively, to the density; see (68), (71).

The domain Ω under consideration is the gearwheel shown in Figure 8. It has 15 teeth and is represented using 180 patches. As Dirichlet data, we choose the restriction of the harmonic function

$$U(x) = \frac{\langle a, x \rangle}{\|x\|^2}, \quad a = (1, 2, 4)$$

to Γ . Then, U is the unique solution of the Dirichlet problem. We discretize the given boundary integral equation by piecewise constant wavelets with three vanishing moments.

In order to measure the error produced by the method, we calculate the approximate solution $U_J = Au_J$ at several points x_i inside the domain, plotted in Figure 8. The last column in the tables below reflects the effect of an a posteriori compression applied to the computed entries of the stiffness matrix. The discrete potentials are denoted by

$$U := [U(x_i)], \quad U_J := [(Au_J)(x_i)]$$

where A stands for the single or double layer operator.

We list in Tables 1 and 2 the results produced by the wavelet Galerkin scheme. For the double-layer approach, the optimal order of convergence of the discrete potential is quadratic with respect to l^∞ -norm over all points x_i .

Table 1. Numerical results with respect to the double layer operator.

| J | N_J | $\ U - U_J\ _\infty$ | cpu-time | a priori (%) | a posteriori (%) |
|-----|--------|----------------------|----------|--------------|------------------|
| 1 | 720 | 4.8e-1 | 1 | 27 | 7.9 |
| 2 | 2880 | 2.7e-1 (1.8) | 10 | 8.7 | 2.3 |
| 3 | 11520 | 7.6e-2 (3.6) | 107 | 3.4 | 0.6 |
| 4 | 46080 | 2.4e-2 (3.1) | 839 | 1.0 | 0.2 |
| 5 | 184320 | 6.0e-3 (4.1) | 4490 | 0.2 | 0.0 |

Table 2. Numerical results with respect to the single layer operator.

| J | N_J | $\ U - U_J\ _\infty$ | cpu-time | a priori (%) | a posteriori (%) |
|-----|--------|----------------------|----------|--------------|------------------|
| 1 | 720 | 4.9e-1 | 1 | 28 | 21 |
| 2 | 2880 | 5.7e-2 (8.6) | 12 | 10 | 7.4 |
| 3 | 11520 | 1.2e-2 (4.5) | 116 | 4.2 | 2.0 |
| 4 | 46080 | 2.8e-3 (4.5) | 1067 | 1.3 | 0.5 |
| 5 | 184320 | 1.0e-3 (2.9) | 6414 | 0.4 | 0.1 |

For the single-layer approach, this order is cubic. But one should mention that one cannot expect the full orders of convergence, due to the reentrant edges resulting from the teeth of the gearwheel.

4.6 Concluding remarks

The above results show how to realize, for any (a priori fixed) level J of resolution, a numerical scheme that solves a boundary integral equation with discretization error accuracy in linear time. As for the quantitative performance, the above examples indicate that accuracy is not degraded at all by the compression and quadrature errors. Moreover, the robustness with respect to the underlying geometry is surprisingly high. The experiences gained in Harbrecht (2001) show that the number of basis functions on the coarsest level may go up to the square root of the overall problem size without spoiling the complexity significantly. Due to the built-in preconditioning, the actual iterative solution of

the linear systems is still by far dominated by the efforts for computing \tilde{A}_J .

The concept is strictly based on a perturbation of the operator but makes no use of adaptivity with respect to the discretization. First tests in this direction have been made in Harbrecht (2002). However, this is somewhat incompatible with the basic structure where all computations are tuned to an a priori fixed highest level J of spatial resolution. Finite subsets of the wavelet basis serve to formulate a Galerkin discretization in the same way as classical settings so that no direct use is made of the full wavelet transformed representation of the boundary integral equation.

Thus, incorporating adaptivity may require an alternative to the entrywise computation of \tilde{A}_J , which we shall comment on later.

There are other ways of accelerating the calculation of matrix/vector products in the above context such as panel clustering (Hackbusch and Nowak, 1989), multipole expansions (Greengard and Rokhlin, 1987), or hierarchical matrices (Hackbusch, 1999). These concepts offer an even better robustness with respect to the geometry since they exploit the smoothness of the integral kernel in \mathbb{R}^d and not of its trace on the manifold. However, these approaches do not allow one to build in preconditioning in such a straightforward way as above and adaptivity is even harder to incorporate. A combination of the different concepts has recently been proposed in Schmidlin, Lage and Schwab (2002), combining the advantages of clustering and wavelet techniques.

5 A NEW ADAPTIVE PARADIGM

So far, we have sketched essentially two different directions where the key features of wavelets listed in Section 2.4 played an essential role. In the context of boundary integral equations, it was established that corresponding operators possess well-conditioned sparse wavelet representations. When dealing with hyperbolic conservation laws, the typical piecewise smooth nature of solutions permits the compression of the flow field based on suitable thresholding strategies applied to multiscale representations of approximate solutions. In both cases, an arbitrary but fixed level of resolution was considered and wavelet concepts were used to precondition or accelerate the numerical processing of the resulting fixed finite dimensional problem.

We shall now turn to recent developments that deviate from this line and aim at combining in a certain sense both effects, namely the sparse representation of functions and the sparse representation of (linear and nonlinear) operators.

The subsequent developments are based on the results in Cohen, Dahmen and DeVore (2001, 2002a,b,c), and Dahlike, Dahmen and Urban (2002).

5.1 Road map

Recall that the classical approach is to utilize a variational formulation of a differential or integral equation mainly as a starting point for the formulation of (Petrov-) Galerkin scheme, which gives rise to a finite dimensional system of linear or nonlinear equations. The finite dimensional problem has then to be solved in an efficient way. As we have seen before, one then faces several obstructions such as ill-conditioning or the instability of the discretizations, for instance, due to the wrong choice of trial spaces. For instance, in the case of the double layer potential operator, stability of the Galerkin scheme is only guaranteed for sufficiently good spatial resolution, that is, sufficient closeness to the infinite dimensional problem. As we shall see below, more severe stability problems are encountered when dealing with noncoercive problems such as saddle point problems. In this case, the trial spaces for the different solution components have to satisfy certain compatibility conditions known as the Ladyženskaja-Babuška-Brezzi (LBB) condition. In brief, although the underlying infinite dimensional problem may be well-posed in the sense of (40), the corresponding finite dimensional problem may not always share this property.

In contrast, we propose here a somewhat different paradigm that tries to exploit the well-posedness of the underlying continuous problem to the best possible extent along the following line:

- (I) Establish well-posedness of the underlying variational problem;
- (II) transform this problem into an equivalent infinite dimensional one which is now well-posed in ℓ_2 ;
- (III) devise a convergent iteration for the infinite dimensional ℓ_2 -problem;
- (IV) only at that stage, realize this iteration approximately with the aid of an adaptive application of the involved (linear or nonlinear) operators.

5.2 The scope of problems — (I)

We describe first the scope of problems we have in mind. We begin with a general format which will then be exemplified by several examples.

For a given (possibly nonlinear) operator \mathcal{F} , the equation

$$\mathcal{F}(u) = f \quad (89)$$

is always understood in a weak sense, namely to find u in some normed space \mathcal{H} such that for given data f

$$(v, \mathcal{F}(u)) = (v, f), \quad \forall v \in \mathcal{H} \quad (90)$$

This makes sense for any $f \in \mathcal{H}'$, the dual of \mathcal{H} (recall (41)) and when \mathcal{F} takes \mathcal{H} onto its dual \mathcal{H}' . In principle, the conservation laws fit into this framework as well. They will be, however, excluded from major parts of the following discussion, since we will assume from now on that \mathcal{H} is a Hilbert space.

The operator \mathcal{F} is often given in a strong form so that the first task is to identify the right space \mathcal{H} for which (90) is well-posed. We have already seen what this means when \mathcal{F} is linear; see (40). The classical Dirichlet problem with $\mathcal{H} = H_0^1(\Omega)$ and the single layer potential equation (67) with $\mathcal{H} = H^{-1/2}(\Gamma)$ are examples. When dealing with nonlinear problems, one may have to be content with corresponding local linearizations are used to define well-posedness. Thus we assume that the Frechét-derivative $D\mathcal{F}(v)$ of \mathcal{F} at v , defined by

$$(z, D\mathcal{F}(v)w) = \lim_{t \rightarrow 0} \frac{1}{t} (z, \mathcal{F}(v + tw) - \mathcal{F}(v)), \quad \forall z \in \mathcal{H}' \quad (91)$$

exists for every v in a neighborhood \mathcal{U} of a solution u of (90) as a mapping from \mathcal{H} onto \mathcal{H}' . In analogy to (40), well-posedness now means that there exist for every $v \in \mathcal{U}$ positive finite constants $c_{\mathcal{F},v}$, $C_{\mathcal{F},v}$ such that

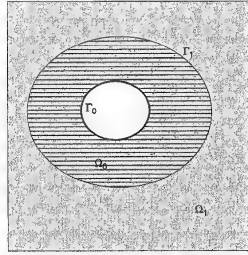
$$c_{\mathcal{F},v} \|w\|_{\mathcal{H}} \leq \|D\mathcal{F}(v)w\|_{\mathcal{H}'} \leq C_{\mathcal{F},v} \|w\|_{\mathcal{H}}, \quad \forall w \in \mathcal{H} \quad (92)$$

We have already seen several examples in Sections 2.5 and 4.1 which, however, are all coercive. We shall, therefore, briefly review several further examples as models for different problem types. They also indicate that, as an additional practical obstruction, the topology for which the problem is well-posed involves norms that are usually difficult to deal with computationally. Most of the examples are actually linear but they may as well play the role of a (local) linearization.

5.2.1 Transmission problem

The following example is interesting because it involves both local and global operators (see Costabel and Stephan, 1990)

$$\begin{aligned} -\nabla \cdot (a \nabla u) &= f && \text{in } \Omega_0, \\ -\Delta u &= 0 && \text{in } \Omega_1, \\ u|_{\Gamma_0} &= 0 \\ \mathcal{H} &:= H_{0,\Gamma_0}^1(\Omega_0) \times H^{-1/2}(\Gamma_1) \end{aligned} \quad \begin{aligned} -v \Delta u + \nabla p &= f && \text{in } \Omega \\ \operatorname{div} u &= 0 && \text{in } \Omega \\ u|_{\Gamma} &= 0 \end{aligned} \quad (93)$$



Both boundary value problems are coupled by the interface conditions:

$$u^- = u^+, \quad (\partial_n u)^- = (\partial_n u)^+$$

A well-posed weak formulation of this problem with respect to the above \mathcal{H} is

$$\begin{aligned} (a \nabla u, \nabla v)_{\Omega_0} + (Wu - (\tfrac{1}{2}I - K')\sigma, v)_{\Gamma_1} &= (f, v)_{\Omega_0}, \\ v &\in H_{0,\Gamma_0}^1(\Omega_0), \\ ((\tfrac{1}{2}I - K')u, \delta)_{\Gamma_1} + (\nabla \sigma, \delta)_{\Gamma_1} &= 0, \quad \delta \in H^{-1/2}(\Gamma_1) \end{aligned}$$

where K, V, W are the double, single layer potential, and hypersingular operator (see Dahmen, Kunoth and Schneider, 2002).

Note that, as an additional obstruction, the occurrence and evaluation of *difficult norms* like $\|\cdot\|_{H^{1/2}(\Gamma)}$, $\|\cdot\|_{H^{-1/2}(\Gamma)}$, $\|\cdot\|_{H^{-1}(\Omega)}$ arises (see Chapter 12, Chapter 13 of this Volume).

5.2.2 Saddle point problems

All the above examples involve coercive bilinear forms. An important class of problems which are no longer coercive are *saddle point problems*. A classical example is

The Stokes System The simplest model for viscous incompressible fluid flow is the Stokes system

where u and p are velocity and pressure respectively (see Brezzi and Fortin, 1991; Girault and Raviart, 1986). The relevant function spaces are

$$\begin{aligned} X &:= (H_0^1(\Omega))^d, \quad M = L_{2,0}(\Omega) \\ &:= \left\{ q \in L_2(\Omega); \int_{\Omega} q = 0 \right\} \end{aligned} \quad (94)$$

In fact, one can show that the range of the divergence operator is $L_{2,0}(\Omega)$. The weak formulation of (93) is

$$\begin{aligned} v(\nabla u, \nabla u)_{L_2(\Omega)} + (\operatorname{div} v, p)_{L_2(\Omega)} &= (f, v), \quad v \in (H_0^1(\Omega))^d \\ (\operatorname{div} u, q)_{L_2(\Omega)} &= 0, \quad q \in L_{2,0}(\Omega) \end{aligned} \quad (95)$$

that is, one seeks a solution (u, p) in the energy space $\mathcal{H} = X \times M = (H_0^1(\Omega))^d \times L_{2,0}(\Omega)$, for which the mapping property (92) can be shown to hold.

First Order Systems One is often more interested in derivatives of the solution of an elliptic boundary value problem which leads to *mixed formulations*. Introducing the fluxes $\theta := -a \nabla u$, (36) can be written as a system of first order equations whose weak formulation reads

$$\begin{aligned} (\theta, \eta) + (\eta, a \nabla u) &= 0, \quad \forall \eta \in (L_2(\Omega))^d, \\ -(\theta, \nabla v) &= (f, v), \quad \forall v \in H_{0,\Gamma_0}^1(\Omega) \end{aligned} \quad (96)$$

One now looks for a solution $(\theta, u) \in \mathcal{H} := (L_2(\Omega))^d \times H_{0,\Gamma_0}^1(\Omega)$. For a detailed discussion in the finite element context, see, for example, Bramble, Lazarov and Pasciak (1997). It turns out that in this case that the Galerkin discretization inherits the stability from the original second order problem.

The General Format The above examples are special cases of the following general problem class. A detailed treatment can be found in Brezzi and Fortin (1991) and Girault and Raviart (1986). Suppose X, M are Hilbert spaces and that $a(\cdot, \cdot)$, $b(\cdot, \cdot)$ are bilinear forms on $X \times X$, $X \times M$ respectively, which are continuous

$$|a(u, v)| \lesssim \|u\|_X \|v\|_X, \quad |b(q, v)| \lesssim \|v\|_X \|q\|_M \quad (97)$$

Given $f_1 \in X'$, $f_2 \in M'$, find $(u, p) \in X \times M := \mathcal{H}$ such that one has for all $(v, q) \in \mathcal{H}$

$$((v, q), \mathcal{F}(u, p)) := \begin{cases} a(u, v) + b(p, v) = (v, f_1), \\ b(q, u) = (q, f_2) \end{cases} \quad (98)$$

Note that when $a(\cdot, \cdot)$ is positive definite symmetric, the solution component u minimizes the quadratic functional

$J(w) := (1/2)a(w, w) - (f_1, w)$ subject to the constraint $b(u, q) = (q, f_2)$, for all $q \in M$, which corresponds to

$$\inf_{w \in \mathcal{H}} \sup_{q \in M} (\tfrac{1}{2}a(w, w) + b(w, q) - (f_1, w) - (q, f_2))$$

This accounts for the term saddle point problem (even under more general assumptions on $a(\cdot, \cdot)$).

In order to write (98) as an operator equation, define the operators A, B by

$$\begin{aligned} a(u, w) &:= (u, Aw), \quad u, w \in X, \\ b(v, p) &:= (Bv, p), \quad p, q \in M \end{aligned}$$

so that (98) becomes

$$\mathcal{F}(u, p) := \begin{pmatrix} A & B' \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} =: f \quad (99)$$

As for the mapping property (92), a simple (sufficient) condition reads as follows (see Brezzi and Fortin, 1991; Girault and Raviart, 1986). If $a(\cdot, \cdot)$ is *elliptic* on

$$\ker B := \{v \in X; b(v, q) = 0, \quad \forall q \in M\}$$

that is,

$$a(v, v) \sim \|v\|_X^2, \quad v \in \ker B \quad (100)$$

and if $b(\cdot, \cdot)$ satisfies the *inf-sup condition*

$$\inf_{q \in M} \sup_{v \in X} \frac{b(v, q)}{\|v\|_X \|q\|_M} > \beta \quad (101)$$

for some positive β , then (97) is well posed in the sense of (92). Condition (101) means that B is surjective (and thus has closed range). Condition (100) is actually too strong. It can be replaced by requiring bijectivity of A on $\ker B$ (see Brezzi and Fortin, 1991).

Aside from leading to large ill-conditioned systems the additional obstructions are the *indefiniteness* of this type of problem and that the well-posedness of the infinite dimensional problem is not automatically inherited by Galerkin discretizations, say. In fact, the trial spaces in X and M have to be compatible in the sense that they satisfy the inf-sup condition (101) *uniformly* with respect to the resolution of the chosen discretizations. This is called the *Ladyženskaya-Babuška-Brezzi condition* (LBB) and may, depending on the problem, be a delicate task (see Chapter 9, this Volume).

5.2.3 A nonlinear model problem

A wide range of phenomena involve the interaction of a (linear) diffusion with a nonlinear reaction or advection

part. We therefore close the list of examples with the simple class of semilinear elliptic boundary value problems. On one hand, it permits a rather complete analysis. On the other hand, it still exhibits essential features that are relevant for a wider scope of nonlinear problems. In this section, we follow Cohen, Dahmen and DeVore (2002b) and suppose that $a(\cdot, \cdot)$ is a continuous bilinear form on a Hilbert space \mathcal{H} endowed with the norm $\|\cdot\|_{\mathcal{H}}$, which is \mathcal{H} -elliptic, that is, there exist positive constants c, C such that

$$c\|v\|_{\mathcal{H}}^2 \leq a(v, v), \quad a(v, w) \leq C\|v\|_{\mathcal{H}}\|w\|_{\mathcal{H}}, \quad \forall v, w \in \mathcal{H} \quad (102)$$

The simplest example is

$$a(v, u) := (\nabla v, \nabla u) + \kappa(v, u), \quad \kappa \geq 0, \quad (v, w) = \int_{\Omega} vw \quad (103)$$

and $\mathcal{H} = H_0^1(\Omega)$ endowed with the norm $\|v\|_{\mathcal{H}} := \|\nabla v\|_{L_2(\Omega)} + \kappa\|v\|_{L_2(\Omega)}$.

Suppose that $\mathcal{G}: \mathbb{R} \rightarrow \mathbb{R}$ is a function with the following property:

P1 the mapping $v \mapsto \mathcal{G}(v)$ takes \mathcal{H} into its dual \mathcal{H}' and is stable in the sense that

$$\|\mathcal{G}(u) - \mathcal{G}(v)\|_{\mathcal{H}'} \leq C(\max\{\|u\|_{\mathcal{H}}, \|v\|_{\mathcal{H}}\})\|u - v\|_{\mathcal{H}}, \quad u, v \in \mathcal{H} \quad (104)$$

where $s \rightarrow C(s)$ is a nondecreasing function of s .

The problem: Given $f \in \mathcal{H}'$ find $u \in \mathcal{H}$ such that

$$(v, \mathcal{F}(u)) := a(v, u) + (v, \mathcal{G}(u)) = (v, f), \quad \forall v \in \mathcal{H} \quad (105)$$

is of the form (90) with $\mathcal{F}(u) = Au + \mathcal{G}(u)$. Note that with the bilinear form from (103), (105) may also arise through an implicit time discretization of a nonlinear parabolic equation.

The unique solvability of (105) is easily ensured when \mathcal{G} is in addition assumed to be *monotone*, that is, $(x - y)(\mathcal{G}(x) - \mathcal{G}(y)) \geq 0$ for $x, y \in \mathbb{R}$. In this case, $\mathcal{F}(u) = f$ is the Euler equation of a convex minimization problem (Cohen, Dahmen and DeVore, 2002b).

A simple example is

$$(v, \mathcal{F}(u)) = \int_{\Omega} \nabla v \nabla u + vu^3 dx, \quad \mathcal{H} = H_0^1(\Omega), \quad \mathcal{H}' = H^{-1}(\Omega) \quad (106)$$

That (at least for $d \leq 3$) $\mathcal{H} = H_0^1(\Omega)$ is indeed the right choice can be seen as follows. The fact that $H^1(\Omega)$ is continuously embedded in $L_4(\Omega)$ for $d = 1, 2, 3$, readily implies that $\mathcal{G}(v) \in H^{-1}(\Omega)$ for $v \in H_0^1(\Omega)$. Moreover,

$$(z, \mathcal{F}(v + tw) - \mathcal{F}(v)) = t(\nabla z, \nabla w) + (z, t^3 \nabla^2 w + t^2 3 \nabla w^2 + t^3 w^3) \text{ so that } (z, D\mathcal{F}(v)w) = (\nabla z, \nabla w) + 3(z, v^2 w) \text{ and hence}$$

$$D\mathcal{F}(v)w = -\Delta w + 3v^2 w \quad (107)$$

Therefore, again by the embedding $H^1 \hookrightarrow L_p$ if $(1/2) < (s/d) + (1/p)$, that is, $\|v\|_{L_4} \lesssim \|v\|_{\mathcal{H}}$ for $d < 4$ with $\mathcal{H} = H_0^1(\Omega)$, we see that

$$\|D\mathcal{F}(v)w\|_{\mathcal{H}'} = \sup_{z \in \mathcal{H}} \frac{(\nabla z, \nabla w) + 3(z, v^2 w)}{\|z\|_{\mathcal{H}}} \lesssim \|w\|_{\mathcal{H}} + \|v\|_{\mathcal{H}}^2 \|w\|_{\mathcal{H}}$$

On the other hand,

$$\|D\mathcal{F}(v)w\|_{\mathcal{H}'} \geq \frac{(\nabla w, \nabla w) + 3(w, v^2 w)}{\|w\|_{\mathcal{H}}} \geq \frac{\|\nabla w\|_{\mathcal{H}}^2}{\|w\|_{\mathcal{H}}} \geq (1 + c(\Omega)^2)^{-1} \|w\|_{\mathcal{H}}$$

where we have used Poincaré's inequality in the last step. Hence we obtain

$$(1 + c(\Omega)^2)^{-1} \|w\|_{\mathcal{H}} \leq \|D\mathcal{F}(v)w\|_{\mathcal{H}'} \lesssim (1 + \|v\|_{\mathcal{H}}^2) \|w\|_{\mathcal{H}}, \quad w \in \mathcal{H} \quad (108)$$

which is (92).

The scope of problems is actually not limited at all to the variational formulation of integral or partial differential equations but covers, for instance, also optimal control problems with PDEs as constraints (see Dahmen and Kunoth, 2002; Kunoth, 2001).

5.3 Transformation into a well-posed ℓ_2 -problem — (II)

Suppose that (90) is well-posed with respect to the energy space \mathcal{H} . In all previous examples, \mathcal{H} was a (closed subspace of a) Sobolev or a product of such spaces. As indicated in Section 2, it is known how to construct wavelet bases for such spaces. In the following, we will therefore assume that Ψ is a Riesz basis for \mathcal{H} .

The transformation of (90) into wavelet coordinates is analogous to the linear case (see Theorem 1 in Section 2.5). In fact, testing in (90) with $v = \psi_{\lambda}$ for all $\lambda \in \mathcal{J}$, defines through $\mathbf{F}(\mathbf{v}) := ((\psi_{\lambda}, \mathcal{F}(v)))_{\lambda \in \mathcal{J}}$ a sequence valued nonlinear mapping \mathbf{F} , which depends on the array $\mathbf{v} \in \ell_2$ via the wavelet expansion $v = \sum_{\lambda \in \mathcal{J}} v_{\lambda} \psi_{\lambda}$. Similarly, the Jacobian of \mathbf{F} at \mathbf{v} acting on $\mathbf{w} \in \ell_2$ is defined by $D\mathbf{F}(\mathbf{v})\mathbf{w} = ((\psi_{\lambda}, D\mathcal{F}(v)w))_{\lambda \in \mathcal{J}}$. Finally, setting as before

$\mathbf{f} := ((\psi_{\lambda}, f))_{\lambda \in \mathcal{J}}$ the following fact can be derived by the same arguments as in Theorem 1.

Theorem 6. Assume that (92) and (30) hold. Then the variational problem (90) is equivalent to $\mathbf{F}(\mathbf{u}) = \mathbf{f}$ where $\mathbf{u} = \sum_{\lambda \in \mathcal{J}} u_{\lambda} \psi_{\lambda}$. Moreover, when the latter problem is well posed in ℓ_2 , that is, for $v = \sum_{\lambda \in \mathcal{J}} v_{\lambda} \psi_{\lambda}$ in some neighborhood \mathcal{U} of the locally unique solution \mathbf{u} of (90)

$$c_0^{-2} c_{\mathcal{F}, \mathbf{v}}^{-1} \|\mathbf{w}\|_{\ell_2} \leq \|D\mathbf{F}(\mathbf{v})\mathbf{w}\|_{\ell_2} \leq C_0^2 C_{\mathcal{F}, \mathbf{v}} \|\mathbf{w}\|_{\ell_2}, \quad \mathbf{w} \in \ell_2 \quad (109)$$

As for the special case (105), note that the monotonicity of \mathcal{G} implies the monotonicity of $\mathbf{G}(\cdot)$, defined by $\mathbf{G}(\mathbf{v}) := ((\psi_{\lambda}, \mathcal{G}(v)))_{\lambda \in \mathcal{J}}$ and hence the positive semidefiniteness of $D\mathbf{G}(\mathbf{v})$; see Cohen, Dahmen and DeVore (2002b) for details.

5.4 An iteration for the infinite dimensional problem — (III)

Once the problem attains a well-conditioned form in ℓ_2 , it makes sense to devise an iterative scheme for the full infinite dimensional problem $\mathbf{F}(\mathbf{u}) = \mathbf{f}$ that converges with a guaranteed error-reduction rate. These iterations will take the form

$$\mathbf{u}^{n+1} = \mathbf{u}^n - C_n(\mathbf{F}(\mathbf{u}^n) - \mathbf{f}), \quad n = 0, 1, 2, \dots \quad (110)$$

where the (infinite) matrix C_n is possibly stage dependent. It can be viewed as a fixed point iteration based on the trivial identity $\mathbf{u} = \mathbf{u} - C(\mathbf{F}(\mathbf{u}) - \mathbf{f})$. We shall indicate several ways of choosing C_n depending on the nature of the underlying variational problem (90).

Gradient Iterations: In the above case of elliptic semilinear problems, the transformed problem still identifies the unique minimum of a convex functional. Thus, it makes sense to consider *gradient iterations*, that is, $C_n = \alpha \mathbf{I}$

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \alpha(\mathbf{F}(\mathbf{u}^n) - \mathbf{f}), \quad n = 0, 1, 2, \dots \quad (111)$$

In fact, one can estimate a suitable positive damping parameter $\alpha > 0$ from the constants in (30), (92), and (104) (see Cohen, Dahmen and DeVore, 2002b for details), so that (111) converges for \mathbf{u}^0 say with a fixed reduction rate $\rho < 1$,

$$\|\mathbf{u}^{n+1} - \mathbf{u}\|_{\ell_2} \leq \rho \|\mathbf{u}^n - \mathbf{u}\|_{\ell_2}, \quad n \in \mathbb{N}_0 \quad (112)$$

For instance, in the linear case $\mathcal{G} \equiv 0$, one can take any $\alpha < 2/(C_0^2 C_A)$ (see (43)) and verify that $\rho = \max\{1 - \alpha c_0^2 C_A, |1 - \alpha C_0^2 C_A|\}$ works.

Least Squares Iteration: Of course, when dealing with indefinite Jacobians, the above scheme will in general no longer work. However, the well-posedness in ℓ_2 offers, in principle, always a remedy which we explain first in the linear case $Au = f$, where this may stand for any of the well-posed problems discussed in Section 5.2. Since then (109) reduces to (43), one can again find a positive α so that $C_n = \alpha A^T$ leads to an iteration

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \alpha A^T(A\mathbf{u}^n - \mathbf{f}), \quad n = 0, 1, 2, \dots \quad (113)$$

that satisfies (112) with some fixed $\rho < 1$. Clearly, this is simply a gradient iteration for the *least squares formulation* $A^T A = A^T \mathbf{f}$ in the wavelet coordinate domain, see also Dahmen, Kunoth and Schneider (2002) for connections with least squares finite element methods.

This is interesting because it suggests an analog also in the *general nonlinear case* even when $D\mathbf{F}(\mathbf{v})$ is indefinite but (92) (or equivalently (109)) holds. In fact, the role of A^T is played by $D\mathbf{F}(\mathbf{u}^n)^T$. Setting

$$\mathbf{R}(\mathbf{v}) := \mathbf{F}(\mathbf{v}) - \mathbf{f} \quad (114)$$

and noting that

$$\mathbf{R}(\mathbf{v}) = \mathbf{F}(\mathbf{v}) - \mathbf{F}(\mathbf{u}) = \left(\int_0^1 D\mathbf{F}(\mathbf{u} + s(\mathbf{v} - \mathbf{u})) ds \right) (\mathbf{v} - \mathbf{u}) \quad (115)$$

one can derive from (109) that the iteration

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \alpha D\mathbf{F}(\mathbf{u}^n)^T \mathbf{R}(\mathbf{u}^n) \quad (116)$$

can be made to satisfy (112) for a suitable positive damping factor α , depending on the constants in (109) and a sufficiently good initial guess \mathbf{u}^0 (which is always needed in the case of only locally unique solutions). Moreover, when being sufficiently close to the solution, $C_n = \alpha D\mathbf{F}(\mathbf{u}^n)^T$ can be frozen to $C = \alpha D\mathbf{F}(\mathbf{u}^*)^T$ so as to still realize a strict error reduction (112); see Cohen, Dahmen and DeVore (2002b) for details.

Uzawa Iteration: Of course, in the above least squares iterations (113) and (116) the damping factor α may have to be chosen rather small, which entails a poor error-reduction rate ρ . Whenever A or the linearization $D\mathbf{F}(\mathbf{v})$ corresponds to a saddle point operator (98) or (99), the squaring of the condition number caused by the least squares formulation can be avoided with the aid of an *Uzawa iteration* (see Dahlke, Dahmen and Urban, 2002; Dahmen, Urban and Vorloper, 2002). We indicate this only for the linear case. Instead of working directly with the wavelet representation $\mathbf{F}(\mathbf{u}, \mathbf{p}) = \mathbf{f}$, one can first eliminate the solution component \mathbf{u} , the velocity in the case of the Stokes problem, on the

infinite dimensional operator level. Recall from (99) that \mathbf{F} takes the form

$$\mathbf{F}(\mathbf{u}, \mathbf{p}) = \begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{pmatrix},$$

$$\mathbf{f}_1 := ((\psi_{X,\lambda}, f_1))_{\lambda \in \mathcal{J}_X}, \quad \mathbf{f}_2 := ((\psi_{M,\lambda}, f_2))_{\lambda \in \mathcal{J}_M} \quad (117)$$

where for given Riesz bases ψ_X, ψ_M of the component spaces of $\mathcal{H} = X \times M$, $\mathbf{A} := a(\psi_X, \psi_X)$, $\mathbf{B} := b(\psi_M, \psi_X)$ are the wavelet representations of the operators A, B in (99), respectively. Recall also that A need only be invertible on the kernel of B . To eliminate \mathbf{u} , one may first have to modify the system as explained in Dahlke, Dahnen and Urban (2002) so as to make the modification of A an automorphism on all of $\ell_2(\mathcal{J}_X)$. Without loss of generality we may therefore assume that this is already the case and that A satisfies (43). Then from the first system $\mathbf{A}\mathbf{u} + \mathbf{B}^T\mathbf{p} = \mathbf{f}_1$, we conclude that $\mathbf{u} = \mathbf{A}^{-1}\mathbf{f}_1 - \mathbf{A}^{-1}\mathbf{B}^T\mathbf{p}$ which, by the second equation in (117), gives $\mathbf{B}\mathbf{u} = \mathbf{B}\mathbf{A}^{-1}\mathbf{f}_1 - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{p} = \mathbf{f}_2$. Hence (117) is equivalent to the Schur complement problem

$$\mathbf{S}\mathbf{p} = \mathbf{B}\mathbf{A}^{-1}\mathbf{f}_1 - \mathbf{f}_2, \quad \mathbf{S} := \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T \quad (118)$$

Once (118) has been solved, the eliminated component can be recovered from (the elliptic problem)

$$\mathbf{A}\mathbf{u} = \mathbf{f}_1 - \mathbf{B}^T\mathbf{p} \quad (119)$$

From the well-posedness of (117) on ℓ_2 , one easily derives that \mathbf{S} is also boundedly invertible on $\ell_2(\mathcal{J}_M)$. So one can find a positive α such that the gradient iteration $\mathbf{p}^{n+1} = \mathbf{p}^n - \alpha(\mathbf{S}\mathbf{p}^n - (\mathbf{B}\mathbf{A}^{-1}\mathbf{f}_1 - \mathbf{f}_2))$ satisfies (112) for some $\rho < 1$. (Actually, the residual may have to be modified again when, as in the case of the Stokes problem, the Lagrange multiplier space M is a subspace of finite codimension in a larger space for which a wavelet Riesz basis is given. For simplicity, we suppress this issue here and refer to Dahlke, Dahnen and Urban (2002) for details. The problem is that the Schur complement is generally not easily accessible, due to the presence of the factor \mathbf{A}^{-1} . However, note that $\mathbf{S}\mathbf{p}^n - (\mathbf{B}\mathbf{A}^{-1}\mathbf{f}_1 - \mathbf{f}_2) = \mathbf{f}_2 - \mathbf{B}\mathbf{u}^n$ whenever \mathbf{u}^n is the solution of the elliptic problem $\mathbf{A}\mathbf{u}^n = \mathbf{f}_1 - \mathbf{B}^T\mathbf{p}^n$. The gradient iteration for the Schur complement problem then takes the form

$$\mathbf{p}^{n+1} = \mathbf{p}^n - \alpha(\mathbf{f}_2 - \mathbf{B}\mathbf{u}^n), \quad \text{where } \mathbf{A}\mathbf{u}^n = \mathbf{f}_1 - \mathbf{B}^T\mathbf{p}^n, \\ n = 0, 1, 2, \dots \quad (120)$$

Thus, the iteration is again of the form we want, but each step requires as an input the solution of a subproblem.

Newton Iteration: Finally, the choice $\mathbf{C}_n := (D\mathbf{F}(\mathbf{u}^n))^{-1}$ in (110) gives rise to the Newton scheme

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \mathbf{w}^n, \quad D\mathbf{F}(\mathbf{u}^n)\mathbf{w}^n = \mathbf{f} - \mathbf{F}(\mathbf{u}^n) = -\mathbf{R}(\mathbf{u}^n) \quad (121)$$

where each step requires the solution of a linear subproblem. While all previous examples have convergence order one, the Newton scheme, in principle, offers even better convergence behavior. We shall address this issue later in more detail.

5.5 Perturbed iteration schemes — (IV)

We have indicated several ways of forming an (idealized) iteration scheme on the full infinite dimensional transformed system $\mathbf{F}(\mathbf{u}) = \mathbf{f}$. We have made essential use of the fact that the transformed system is well posed in ℓ_2 in the sense of (109). Recall that this hinges on the mapping property (92) induced by the original continuous problem and the availability of a Riesz basis ψ for the corresponding energy space \mathcal{H} , (30). The final step is to realize the idealized iteration numerically. We shall do so *not* by choosing some fixed finite dimensional space on which the problem is projected, as was done in previous sections, but rather by approximating at each step the true residual $\mathbf{r}^n := \mathbf{C}_n(\mathbf{F}(\mathbf{u}^n) - \mathbf{f}) = \mathbf{C}_n\mathbf{R}(\mathbf{u}^n)$ within a suitable dynamically updated accuracy tolerance. Again we wish to avoid choosing for this approximation any a priori, fixed, finite dimensional space but try to realize the required accuracy at the expense of possibly few degrees of freedom. The whole task can then be split into two steps:

- Assuming that, in each case at hand, a computational scheme is available that allows us to approximate the residuals within each desired target accuracy, determine first for which dynamic tolerances the iteration will converge in the sense that for any given target accuracy ϵ , the perturbed iteration outputs a finitely supported vector $\mathbf{u}(\epsilon)$ such that $\|\mathbf{u} - \mathbf{u}(\epsilon)\|_{\ell_2} \leq \epsilon$. Thus on the numerical side, all approximations will take place in the Euclidean metric. Note however that because of (30), this implies that

$$\left\| \mathbf{u} - \sum_{\lambda \in \text{supp } \mathbf{u}(\epsilon)} \langle \mathbf{u}(\epsilon), \psi_\lambda \rangle \psi_\lambda \right\|_{\mathcal{H}} \leq C_\psi \epsilon \quad (122)$$

- Once this has been clarified, one has to come up with concrete realizations of the residual approximations appearing in the idealized iteration. It is clear from the above examples that the crucial task is to approximate with possibly few terms the sequence $\mathbf{F}(\mathbf{u}^n)$ in ℓ_2 .

Moreover, we will have to make precise what we mean by 'possibly few terms', that is, we have to analyze the computational complexity of the numerical schemes.

We shall address first (a) under the assumption that we are given a numerical scheme

RES $[\eta, \mathbf{C}, \mathbf{F}, \mathbf{f}, \mathbf{v}] \rightarrow \mathbf{r}_\eta$: WHICH FOR ANY POSITIVE TOLERANCE η AND ANY FINITELY SUPPORTED INPUT \mathbf{v} OUTPUTS A FINITELY SUPPORTED VECTOR \mathbf{r}_η SATISFYING

$$\|\mathbf{C}\mathbf{R}(\mathbf{v}) - \mathbf{r}_\eta\|_{\ell_2} \leq \eta \quad (123)$$

The second ingredient is a routine

COARSE $[\eta, \mathbf{w}] \rightarrow \mathbf{w}_\eta$: WHICH FOR ANY POSITIVE TOLERANCE η AND ANY FINITELY SUPPORTED INPUT VECTOR \mathbf{w} PRODUCES A FINITELY SUPPORTED OUTPUT \mathbf{w}_η WITH POSSIBLY FEW ENTRIES (SUBJECT TO CONSTRAINTS THAT WILL BE SPECIFIED LATER) SUCH THAT

$$\|\mathbf{w} - \mathbf{w}_\eta\|_{\ell_2} \leq \eta \quad (124)$$

This routine will be essential later to control the complexity of the overall perturbed iteration scheme.

Next we need some initialization, that is, an initial guess \mathbf{u}^0 and an error bound

$$\|\mathbf{u} - \mathbf{u}^0\|_{\ell_2} \leq \epsilon_0 \quad (125)$$

In general, such an estimate depends on the problem at hand. In the case of semi-linear elliptic problems, such a bound is, for instance, $\epsilon_0 = c_\psi^{-1} c_X^{-1} (\|\mathbf{f}\|_{\ell_2} + \|\mathbf{f}\|_{\ell_2})$ for $\mathbf{u}^0 = \mathbf{0}$; see Cohen, Dahnen and DeVore (2002b) for details.

As a final prerequisite, one can find, again as a consequence of (109), for each of the above iteration schemes, (110) a constant β such that

$$\|\mathbf{u} - \mathbf{v}\|_{\ell_2} \leq \beta \|\mathbf{C}\mathbf{R}(\mathbf{v})\|_{\ell_2} \quad (126)$$

holds in a neighborhood of the solution \mathbf{u} . The perturbed iteration scheme may now be formulated as follows.

SOLVE $[\epsilon, \mathbf{C}, \mathbf{R}, \mathbf{u}^0] \rightarrow \tilde{\mathbf{u}}(\epsilon)$

- CHOOSE SOME $\tilde{\rho} \in (0, 1)$. SET $\tilde{\mathbf{u}}^0 = \mathbf{u}^0$, THE CORRESPONDING INITIAL BOUND ϵ_0 ACCORDING TO THE ABOVE INITIALIZATION, AND $j = 0$.
- IF $\epsilon_j \leq \epsilon$ STOP AND OUTPUT $\tilde{\mathbf{u}}(\epsilon) := \tilde{\mathbf{u}}^j$; ELSE SET $\mathbf{v}^0 := \tilde{\mathbf{u}}^j$ AND $k = 0$

(II.1) SET $\eta_k := \tilde{\rho}^k \epsilon_j$ AND COMPUTE

$$\mathbf{r}^k = \text{RES}[\eta_k, \mathbf{C}, \mathbf{F}, \mathbf{f}, \mathbf{v}^k], \quad \mathbf{v}^{k+1} = \mathbf{v}^k - \mathbf{r}^k$$

(II.2) IF

$$\beta(\eta_k + \|\mathbf{r}^k\|_{\ell_2}) \leq \frac{\epsilon_j}{2(1 + 2C^*)} \quad (127)$$

SET $\tilde{\mathbf{v}} := \mathbf{v}^k$ AND GO TO (III). ELSE SET $k + 1 \rightarrow k$ AND GO TO (II.1).

(III) COARSE $[(2C^*\epsilon_j)/(2(1 + 2C^*))], \tilde{\mathbf{v}}] \rightarrow \tilde{\mathbf{u}}^{j+1}$, $\epsilon_{j+1} = \epsilon_j/2$, $j + 1 \rightarrow j$, GO TO (II)

The constant C^* depends on the particular realization of the routine COARSE and will be specified later.

Note that step (III) is just an approximation of the updates in (110). This applies until the stopping criterion (127) is met. This is a posteriori information based on the numerical residual. The fact that there is actually a uniform bound K for the number of updates in step (III), independent of the data and the target accuracy, until (127) is satisfied and a coarsening step (III) is carried out, relies on (126) and the underlying well-posedness (109). The parameter $\tilde{\rho}$ is actually allowed here to be smaller than the true error-reduction rate ρ in (112) for which only a poor or a possibly too pessimistic estimate may be available.

One can show by fairly straightforward perturbation arguments that the choice of accuracy tolerances in SOLVE implies convergence (Cohen, Dahnen and DeVore, 2002b).

Proposition 1. The iterates $\tilde{\mathbf{u}}^j$ produced by the scheme SOLVE satisfy

$$\|\mathbf{u} - \tilde{\mathbf{u}}^j\|_{\ell_2} \leq \epsilon_j \quad (128)$$

so that in particular $\|\mathbf{u} - \tilde{\mathbf{u}}(\epsilon)\|_{\ell_2} \leq \epsilon$. By (30), this means

$$\left\| \mathbf{u} - \sum_{\lambda \in \Lambda(\epsilon)} \langle \tilde{\mathbf{u}}(\epsilon), \psi_\lambda \rangle \psi_\lambda \right\|_{\mathcal{H}} \leq C_\psi \epsilon \quad (129)$$

where C_ψ is the constant from (30) and $\Lambda(\epsilon) := \text{supp } \tilde{\mathbf{u}}(\epsilon)$.

6 CONSTRUCTION OF RESIDUAL APPROXIMATIONS AND COMPLEXITY ANALYSIS

It remains to construct concrete realizations of the routines RES and COARSE. It turns out that the development of such routines is closely intertwined with their complexity analysis. Since the conceptual tools are probably unfamiliar in the context of numerical simulation, we highlight some of them in the next section.

6.1 Best N -term approximation

A lower bound for the computational complexity of Solve is, of course, the growth of the supports of outputs in step (II), which determines how the overall number of degrees of freedom grows until the target accuracy is reached. Therefore, a lower bound for the computational complexity of Solve is given by the number of terms needed to recover the true solution u in ℓ_2 within accuracy ϵ . This is the issue of best N -term approximation in ℓ_2 . Thus the question arises whether, or under which circumstances, Solve can actually attain this lower bound, at least asymptotically. Since best N -term approximation limits what can be achieved at best, we briefly review some relevant features concerning best N -term approximation.

There are two intimately connected ways of looking at an error analysis for N -term approximation. In the first, we can specify the target accuracy ϵ and ask what is the smallest number $N(\epsilon)$ of terms needed to recover a given object? The second view is to assume we are given a budget N of terms and ask what accuracy $\epsilon(N)$ can be achieved with the best selection of N terms? The process of selecting such N terms is obviously nonlinear. This can be formalized by defining the following error of approximation:

$$\sigma_{N,\ell_2}(v) := \inf_{\substack{\text{supp } w \leq N}} \|v - w\|_{\ell_2} \quad (130)$$

Obviously, $\sigma_{N,\ell_2}(v)$ is attained by $w = v_N$ comprised of the N largest terms of v in modulus. Note that this is not necessarily unique since several terms may have equal modulus. Analogously one can define $\sigma_{N,\mathcal{H}}(v)$ by looking for the best approximation of v in \mathcal{H} by a linear combination of at most N wavelets. One easily infers from (30) that

$$c_\Psi \sigma_{N,\ell_2}(v) \leq \sigma_{N,\mathcal{H}}(v) \leq C_\Psi \sigma_{N,\ell_2}(v) \quad (131)$$

Best N -term approximations in the Euclidean metric yield therefore near-best N -term approximations in \mathcal{H} . Hence, an element in \mathcal{H} can be approximated well with relatively few terms if and only if this is true for its coefficient array in ℓ_2 .

We shall proceed with identifying classes of sequences in ℓ_2 for which $\epsilon(N)$ decays like N^{-s} , since these are the rates that can be expected from approximations based on spatial refinements (h -methods). To this end, consider the classes

$$\begin{aligned} \mathcal{A}^s(\mathcal{H}) &:= \{v \in \mathcal{H} : \sigma_{N,\mathcal{H}}(v) \lesssim N^{-s}\} \\ \mathcal{A}^s(\ell_2) &:= \{v \in \ell_2 : \sigma_{N,\ell_2}(v) \lesssim N^{-s}\} \end{aligned} \quad (132)$$

These are normed linear spaces endowed with the norms

$$\|v\|_{\mathcal{A}^s(\mathcal{H})} := \sup_{N \in \mathbb{N}} N^s \sigma_{N,\mathcal{H}}(v) \quad \|v\|_{\mathcal{A}^s(\ell_2)} := \sup_{N \in \mathbb{N}} N^s \sigma_{N,\ell_2}(v)$$

Thus to achieve a target accuracy ϵ , the order of $N(\epsilon) \sim \epsilon^{-1/s}$ terms are needed for $v \in \mathcal{A}^s(\mathcal{H})$ or $v \in \mathcal{A}^s(\ell_2)$. Hence the larger $s > 0$ the less terms suffice.

Which property makes a function v or its coefficient sequence v sparse in the above sense is best explained when \mathcal{H} is a Sobolev space $\mathcal{H} = H^s$ over some domain in \mathbb{R}^d . One can then show that for any positive δ (cf. DeVore, 1998; Bergh and Löfström, 1976; Cohen, 2000)

$$\begin{aligned} \ell_q \subset \mathcal{A}^s(\ell_2) \subset \ell_{q+\delta}, \quad B_q^{s+\delta}(L_q) \subset \mathcal{A}^s(H^s) \\ \subset B_q^{s+\delta-\delta}(L_q), \quad \frac{1}{q} = s + \frac{1}{2} \end{aligned} \quad (133)$$

Here, ℓ_q consists of the q -summable sequences, while $B_q^s(L_q)$ denotes a Besov space consisting roughly of functions with smoothness s measured in L_q (see DeVore, 1998; DeVore and Lorentz, 1993; Bergh and Löfström, 1976; Cohen, 2003; Barinka, Dahlke and Dahmen, 2003; Barinka et al., 2001; Berger and Olliger, 1984 for precise definitions). For a certain smoothness range, depending on the regularity of the wavelet basis these spaces can also be characterized through wavelet bases. In fact, for $0 < q \leq \infty$ one has

$$\|v\|_{B_q^s(L_q)}^q \sim \|v\|_{L_q}^q + \sum_{\lambda \in \mathcal{J}} 2^{sq|\lambda|} \|(v, \tilde{\psi}_\lambda)_\psi\|_{L_q}^q \quad (134)$$

which, due to the equivalence $H^s = B_2^s(L_2)$, covers (32) as a special case (see e.g. Cohen, 2000, 2003; DeVore, 1998). It is important to note here that the smaller the q , the weaker is the smoothness measure. By the Sobolev embedding theorem, the value of q given by (133) gives the weakest possible measure so that smoothness of order $sd + \alpha$ in L_q guarantees Sobolev regularity of order α corresponding to the anchor space $\mathcal{H} = H^s$ (a Sobolev space of order α or a closed subspace defined, for example, by homogeneous boundary conditions). This is illustrated in Figure 9 below. Each point in the $(1/q, s)$ -plane corresponds to a smoothness space (actually to a class of smoothness spaces) describing smoothness s measured in L_q . In our case, we have $X = H^s$ and $p = 2$. The spaces located left of the line with slope d emanating from X are embedded in X . The spaces of smoothness $\alpha + sd$ on the vertical line above X are essentially those whose elements can be approximated with accuracy $O(N^{-\alpha})$ by approximants from quasi-uniform meshes, that is, with equidistributed degrees of freedom. In the present terms, this means just keeping all wavelets up to some scale J , say (or equivalently working with uniform meshes), so that $N \sim 2^{Jd}$ would require the function v to belong to $B_{\infty}^{s+sd}(L_2)$, which is very close to the Sobolev space H^{s+sd} . The spaces on the critical embedding line, however, are characterized by nonlinear approximation like best N -term approximation.

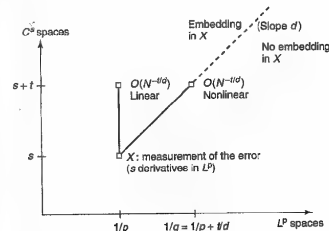


Figure 9. Topography of smoothness spaces.

Thus, while the spaces obtained when moving to the right, away from the vertical line on the same smoothness level, grow and admit increasingly stronger singularities. This loss of regularity can be compensated by judiciously placing the degrees of freedom so as to retain the same convergence rates in terms of degrees of freedom N . Since H^{s+sd} is a much smaller space than $B_q^{s+sd}(L_q)$ this indicates the possible gain offered by nonlinear approximation schemes like best N -term approximation over simpler schemes based on a priori fixed discretizations.

Of course, it remains to see whether this potential can be exploited by adaptive schemes.

Tree Structures: The above notion of best N -term approximation puts no constraints on the distribution of significant coefficients. In the context of conservation laws, it was important that the significant coefficients were arranged in tree-like structures which correspond to local mesh refinements. Thus interrelating the selection of wavelets with locally refined meshes is one reason for imposing some sort of constraint on the distribution of wavelet coefficients. Another reason arises when approximating the quantities $F(v)$ when v is some finitely supported vector. Intuitively, one might expect that the nonlinearity in F makes the effect of a term v_λ with large $|\lambda|$ cascade down to lower levels in a neighborhood of the support S_λ , which also gives rise to tree-like structures.

Let us first explain what we mean by a tree structure associated to the set of wavelet indices. In the simplest case of a one dimensional basis $\psi_k = \psi_{j,k} = 2^{j/2} \psi(2^j \cdot - k)$, this structure is obvious: each index (j, k) has two children $(j+1, 2k)$ and $(j+1, 2k+1)$. A similar tree structure can be associated to all available constructions of wavelet bases on a multidimensional domain: to each index λ , one can be assign $m(\lambda) \geq 2$ children μ such that $|\mu| = |\lambda| + 1$, where

$m(\lambda)$ might vary from one index to another but is uniformly bounded by some fixed K . We shall use the notation $\mu < \lambda$ ($\mu \leq \lambda$) in order to express that μ is a descendent of λ (or equals λ) in the tree. We also have the property

$$\mu < \lambda \implies S_\mu \subset S_\lambda \quad (135)$$

where we recall that $S_\lambda := \text{supp } \psi_\lambda$. A set $T \subset \mathcal{J}$ is called a tree if $\lambda \in T$ implies $\mu \in T$ whenever $\lambda < \mu$.

If the tree $T \subset \mathcal{J}$ is finite, we define the set $\mathcal{L} = \mathcal{L}(T)$ of outer leaves as the set of those indices outside the tree whose parent belongs to the tree

$$\mathcal{L} := \{\lambda \in \mathcal{J} : \lambda \notin T, \lambda < \mu \implies \mu \in T\} \quad (136)$$

The set $\mathcal{L}(T)$ plays the role of a (locally refined) mesh. In fact, one readily confirms that

$$\|v - v|_T\|_{\ell_2}^2 = \sum_{\lambda \in \mathcal{L}} |v_\lambda|^2 = \sum_{\lambda \in \mathcal{L}(T)} \sum_{\mu \leq \lambda} |v_\mu|^2 \quad (137)$$

which suggests considering the quantities

$$\tilde{v}_\lambda := \sum_{\mu \leq \lambda} |v_\mu|^2 \quad (138)$$

These quantities measure in some sense a local error associated with the spatial location of ψ_λ . To see this, suppose that the wavelets have the form $\psi_\lambda = \omega_\lambda \theta_\lambda$ where ω_λ are some positive weights (see (33)) and Θ is a Riesz basis for L_2 (which is the case for all constructions considered in Section 2). Then, by (30)

$$\left\| \sum_{\lambda \in \mathcal{J}} v_\lambda \psi_\lambda \right\|_{\mathcal{H}} \sim \|v\|_{\ell_2} \sim \left\| \sum_{\lambda \in \mathcal{J}} v_\lambda \theta_\lambda \right\|_{L_2}$$

so that

$$\left\| v - \sum_{\lambda \in T} v_\lambda \psi_\lambda \right\|_{\mathcal{H}} \sim \left\| \sum_{\lambda \in \mathcal{L}} v_\lambda \theta_\lambda \right\|_{L_2} \quad (139)$$

Note that the right hand side can be localized. In fact, for $\mu \in \mathcal{L}(T)$

$$\begin{aligned} \left\| \sum_{\lambda \in \mathcal{L}} v_\lambda \theta_\lambda \right\|_{L_2(S_\mu)}^2 &= \left\| \sum_{|\mu| \leq |\lambda| : S_\lambda \cap S_\mu \neq \emptyset} v_\lambda \theta_\lambda \right\|_{L_2(S_\mu)}^2 \\ &\lesssim \sum_{|\mu| \leq |\lambda| : S_\lambda \cap S_\mu \neq \emptyset} v_\lambda^2 \lesssim \sum_{\lambda \in \mathcal{L}(T), S_\lambda \cap S_\mu \neq \emptyset} \tilde{v}_\lambda^2 \end{aligned} \quad (140)$$

It has been shown in Cohen, Dahmen and DeVore (2002c) that any tree T can be expanded to a tree \tilde{T} such that $\#\tilde{T} \lesssim \#T$ but for any $\mu \in \mathcal{L}(T)$ only for a uniformly

bounded finite number of $\lambda \in \mathcal{L}(\tilde{T})$ one has $S_\lambda \cap S_\lambda \neq \emptyset$. Hence a finite number of the terms \tilde{v}_λ bound the local error on S_λ .

A natural idea for constructing 'good' meshes – or equivalently 'good trees' identifying spans of wavelets – is to *equilibrate* these local errors. However, it turns out that this will not necessarily minimize the error $\|v - v_{T_N}\|_{\ell_2}$ over all trees of a fixed cardinality $N = \#(T)$ (see Cohen, Dahmen and DeVore, 2002c). To formalize this, we define an error for N -term tree approximation that is the exact *tree analog* of the best (unconstrained) N -term approximation defined in (130).

$$\sigma_{N,\ell_2}^{\text{tree}}(v) := \min \{\|v - w\|_{\ell_2} : T := \text{supp } w \text{ is a tree and } \#T \leq N\} \quad (141)$$

Any minimizing tree will be denoted by $T_N(v)$. We define now in analogy to (131) the sequence space

$$\mathcal{A}_{\text{tree}}^s(\ell_2) := \{v \in \ell_2 : \sigma_{N,\ell_2}^{\text{tree}}(v) \lesssim N^{-s}\} \quad (142)$$

endowed with the quasi-norm

$$\|v\|_{\mathcal{A}_{\text{tree}}^s(\ell_2)} := \sup_{N \in \mathbb{N}} N^s \sigma_{N,\ell_2}^{\text{tree}}(v) \quad (143)$$

Analogously, we can define the counterpart $\mathcal{A}_{\text{tree}}^s(\mathcal{H})$ in \mathcal{H} . As in (131) the error of tree approximation of $v \in \mathcal{A}_{\text{tree}}^s(\ell_2)$ decays like the error of the corresponding tree approximations to v in \mathcal{H} .

In spite of the conceptual similarity, there is an important difference between best tree and best unconstrained N -term approximation. At least for any finitely supported v , the latter one is easily determined by (quasi-) sorting by size thresholding. Determining the best tree however, is much harder. However, since one obtains a near-best approximation in the energy norm anyway, we can be content with *near-best* tree approximation in ℓ_2 as well. More precisely, given a fixed constant $C^* \geq 1$, a tree $T(\eta, v)$ is called an (η, C^*) -near best tree for v if $\|v - v_{T(\eta, v)}\|_{\ell_2} \leq \eta$ and whenever any other tree T satisfies $\|v - v_T\|_{\ell_2} \leq \eta/C^*$ one has $\#(T(\eta, v)) \leq C^*\#(T)$. It is remarkable that, according to Binev and DeVore (2002), such near-best trees can be constructed in *linear time*. This can be achieved with the aid of modified thresholding strategies working in the present setting with the quantities \tilde{v}_λ^2 (138) as local error functionals. We shall invoke this method to construct near-best trees.

Since the selections of terms are constrained by the imposed tree structure, one always has

$$\sigma_{\#T(\eta, \ell_2)}(v) \leq \|v - v_{T(\eta, v)}\|_{\ell_2} \quad (144)$$

However, for a wide class of functions in \mathcal{H} , one actually does not lose too much with regard to an optimal work/accuracy rate. To explain this, we consider again the above scenario $\mathcal{H} = H^t$. The following fact has been shown in Cohen, Dahmen and DeVore (2002c).

Remark 4. For $\mathcal{H} = H^t$ one has $B_q^{t+sd}(L_q) \hookrightarrow \mathcal{A}_{\text{tree}}^s(\mathcal{H})$ whenever $q^{-1} < s + 1/2$.

Thus, as soon as the smoothness space is strictly left of the Sobolev embedding line, its elements have errors of tree approximations that decay like $(\#T_N(v))^{-s}$; see Figure 9. Moreover, this rate is known to be sharp, that is,

$$\sup_{\|v\|_{B_q^{t+sd}(L_q)}=1} \inf_N N^s \sigma_{N,H^t}(v) \gtrsim 1 \quad (145)$$

which means that on the class $B_q^{t+sd}(L_q)$, under the above restriction of q , tree approximations give the same asymptotic error decay as best N -term approximations. The smaller the discrepancy $\delta := s + (1/2) - (1/q) > 0$, the larger the space $B_q^{t+sd}(L_q)$ admitting stronger singularities. In fact, when $\sup\{s : u \in H^{t+sd}\}$ is strictly smaller than $\sup\{s : u \in B_q^{t+sd}(L_q)\}$ the asymptotic work/accuracy rate achieved by meshes corresponding to the trees $T_N(\cdot)$ is strictly better than that for uniform mesh refinements. This is known to be the case, for instance, for solutions u of elliptic boundary value problems on Lipschitz domains when δ is sufficiently small (see Dahlike and DeVore, 1997; Dahlike, 1999).

Thus, the question that guides the subsequent discussion can be formulated as follows: *Can one devise the routines RS and COARSE in such a way that the computational work and storage needed to produce the output $u(\epsilon)$ of SOLVE, stays proportional to $\epsilon^{-1/2}$, uniformly in ϵ , whenever the unknown solution u belongs to $\mathcal{A}_{\text{tree}}^s(\mathcal{H})$, or even to $\mathcal{A}^s(\mathcal{H})$?*

6.2 Realization of residual approximations

We shall always assume that we have full access to the given data f . Depending on some target accuracy, one should therefore think of f as a finite array that approximates some 'ideal' data accurately enough. Moreover, these data are (quasi-)ordered by their modulus. Such a quasi-ordering, based on binary binning can be realized in linear time (see e.g. Barinka, 2003). In particular, this allows us to obtain coarser approximations f_n , satisfying $\|f - f_n\|_{\ell_2} \leq \eta$ with the aid of the simplest version of the routine COARSE, realized by adding $\|f_n\|^2$ in the direction of increasing size until the sum exceeds η^2 ; see Cohen, Dahmen and DeVore (2001) for details. As a central task, one further

has to approximate the sequence $F(v)$ for any given finitely supported input v that we shall now describe.

Linear Operators: It will be instructive to consider first the linear case $F(v) = Av$ when A is the wavelet representation of the underlying operator. We shall describe an algorithm for the fast computation of Av . So far, the only property of A that we have used is the norm equivalence (30). Now the cancellation properties (26) come into play. We have seen in Section 4.2 that they imply the quasi-sparsity of a wide class of linear operators. The relevant notion can be formulated as follows (Cohen, Dahmen and DeVore, 2001). A matrix C is said to be s^* -compressible – $C \in C_{s^*}$ – if for any $0 < s < s^*$ and every $j \in \mathbb{N}$ there exists a matrix C_j with the following properties: For some summable sequence, $(\alpha_j)_{j=1}^\infty$ ($\sum_j \alpha_j < \infty$) C_j is obtained by replacing all but the order of $\alpha_j 2^j$ entries per row and column in C by zero and satisfies

$$\|C - C_j\| \leq C \alpha_j 2^{-js^*}, \quad j \in \mathbb{N} \quad (146)$$

Specifically, wavelet representations of differential and also the singular integral operators from Sections 4.2 and 4.1 fall into this category for values of s^* , that depend, in particular, on the regularity of the wavelets (see Cohen, Dahmen and DeVore, 2001; Dahlike, Dahmen and Urban, 2002; Stevenson, 2003).

In order to describe the essence of an approximate application scheme for compressible matrices, we abbreviate for any finitely supported v the best 2^j -term approximations by $v_{[j]} := v_{2^j}$ ($v_{[-1]} \equiv \emptyset$) and define

$$w_j := A_j v_{[0]} + A_{j-1}(v_{[1]} - v_{[0]}) + \dots + A_0(v_{[j]} - v_{[j-1]}) \quad (147)$$

as an approximation to Av . Obviously this scheme is *adaptive* in that it exploits directly information on v . In fact, if $A \in C_{s^*}$, then the triangle inequality together with the above compression estimates yield for any fixed $s < s^*$

$$\|Av - w_j\|_{\ell_2} \leq C \left(\|v - v_{[j]}\|_{\ell_2} + \sum_{i=0}^j \alpha_i 2^{-is^*} \|v_{[i]} - v_{[i-1]}\|_{\ell_2} \right) \lesssim \sigma_{2^j, \ell_2}(v) \quad (148)$$

One can now exploit the a posteriori information offered by the quantities $\sigma_{2^j, \ell_2}(v)$ to choose the smallest j for which the right hand side of (148) is smaller than a given target accuracy η and set $w_\eta := w_j$. Since the sum is finite for each finitely supported input v such a j does indeed exist. This leads to a concrete multiplication scheme (see

Cohen, Dahmen and DeVore, 2001; Barinka *et al.*, 2001 for a detailed description, analysis and implementation), which we summarize as follows:

APPLY $[\eta, A, v] \rightarrow w_\eta$; DETERMINES FOR ANY FINITELY SUPPORTED INPUT v A FINITELY SUPPORTED OUTPUT w_η SUCH THAT

$$\|Av - w_\eta\|_{\ell_2} \leq \eta \quad (149)$$

Depending on the compressibility range s^* this scheme can be shown to exhibit the same work/accuracy rate as the best (unconstrained) N -term approximation in ℓ_2 as stated by the following result (Cohen, Dahmen and DeVore, 2001).

Theorem 7. Suppose that $A \in C_{s^*}$ and that for some $0 < s < s^*$, $v \in \mathcal{A}^s(\ell_2)$. Then, Av is also in $\mathcal{A}^s(\ell_2)$. Moreover, for any finitely supported v the output $w_\eta = \text{APPLY}(\eta, C, v)$ satisfies:

- (i) $\|w_\eta\|_{\mathcal{A}^s(\ell_2)} \lesssim \|v\|_{\mathcal{A}^s(\ell_2)}$;
- (ii) $\#\text{supp } w_\eta \lesssim \|v\|_{\mathcal{A}^s(\ell_2)}^{1/s} \eta^{-1/s}$, $\#\text{flops} \lesssim \#\text{supp } v + \|v\|_{\mathcal{A}^s(\ell_2)} \eta^{-1/s}$,

where the constants in these estimates depend only on s when s is small.

The above work count is based on the tacit assumption that the entries of A can be computed with sufficient accuracy on average at unit cost. This can be verified for constant coefficient differential operators and spline wavelets. In general, the justification of such an assumption is less clear. We shall return to this point later.

The Nonlinear Case — Prediction of Significant Coefficients: In this case, the point of view changes somewhat. The question to be addressed first is the following:

Given any $\eta > 0$ and an (η, C^*) -near best tree $T(\eta, v)$ of v , find a possibly small tree T_η such that for some constant C

$$T^*(C\eta, F(v)) \leq T_\eta \quad (150)$$

where $T^*(C\eta, F(v))$ is a smallest tree realizing accuracy $C\eta$.

Thus, we are asking for quantitative estimates concerning the effect of a nonlinearity on contributions with different length scales, a question of central importance in several areas of applications such as turbulence analysis. Using trees now already anticipates the need for taking the (quasi-) effect of higher frequency on lower ones into account.

Of course, tight estimates of that sort must incorporate some knowledge about the character of the nonlinearity.

Nonlinearities of at most *power growth* have been studied recently in Cohen, Dahmen and DeVore (2002c) and we briefly review some of the main findings. For instance, when the operator involves a local composition operator G as in (105) 'power growth' means that for some $p > 0$ $|G^{(k)}(x)| \lesssim (1+|x|)^{(p-k)_+}$. In fact, one can show that for $\mathcal{H} = H^t$ (on some domain of spatial dimension d) one has $\mathcal{G}: \mathcal{H} \rightarrow \mathcal{H}'$ provided that

$$p < p^* := \frac{d+2t}{d-2t} \text{ when } t < \frac{d}{2}, \quad p > 0 \text{ when } t \geq \frac{d}{2} \quad (151)$$

(see Cohen, Dahmen and DeVore, 2002c). The analysis in Cohen, Dahmen and DeVore (2002c) covers a much wider class of nonlinear operators including those that depend on several components involving also derivatives of the arguments $G(D^{\alpha_1}v_1, \dots, D^{\alpha_m}v_m)$. For instance, the convective term in the Navier Stokes equations is covered. In order to convey the main ideas while keeping the exposition as simple as possible, we confine the subsequent discussion to the above special situation. Using the locality of the nonlinearity, the cancellation properties of the wavelets as well as certain norm equivalences for Besov spaces in terms of weighted sequence norms for the wavelet coefficients, one can derive estimates of the form

$$|\mathbf{F}(v)_\lambda| \lesssim \sup_{S_\lambda \cap S_\mu \neq \emptyset} |v_\mu| 2^{-\gamma(|\lambda| - |\mu|)} \quad (152)$$

where for $\mathcal{H} = H^t$ a typical value for γ is $\gamma = t + \tilde{m} + d/2$. It measures in some sense the compressibility of the nonlinear map.

How to predict good trees for $\mathbf{F}(v)$ from those for v for the above-mentioned type of nonlinearities can be sketched as follows (cf. Cohen, Dahmen and DeVore, 2002c). For a given target accuracy ϵ , $j = 0, 1, \dots$ and a given v , we consider the near-best trees

$$\mathcal{T}_j := \mathcal{T}\left(\frac{2^j \epsilon}{1+j}, v\right) \quad (153)$$

and the corresponding expanded trees $\tilde{\mathcal{T}}_j$ mentioned before. By construction, these trees are nested in the sense that $\tilde{\mathcal{T}}_j \subset \tilde{\mathcal{T}}_{j-1}$. We shall use the difference sets

$$\Delta_j := \tilde{\mathcal{T}}_j \setminus \tilde{\mathcal{T}}_{j+1} \quad (154)$$

in order to build a tree, which will be adapted to $w = \mathbf{F}(v)$. They represent the 'energy' in v reflecting the next higher level of accuracy. Now we introduce the parameter

$$\alpha := \frac{2}{2\gamma - d} > 0 \quad (155)$$

where γ is the constant in (152) and for each $\mu \in \Delta_j$, we define the *influence set*

$$\Lambda_{\epsilon, \mu} := \{\lambda: S_\lambda \cap S_\mu \neq \emptyset \text{ and } |\lambda| \leq |\mu| + \alpha j\} \quad (156)$$

Thus the amount αj by which the level $|\mu|$ is exceeded in $\Lambda_{\epsilon, \mu}$ depends on the 'strength' of v_μ expressed by the fact that $\mu \in \Delta_j$. We then define \mathcal{T}_ϵ as the union of these influence sets

$$\mathcal{T}_\epsilon := \bigcup_{\mu \in \tilde{\mathcal{T}}_0} \Lambda_{\epsilon, \mu} \quad (157)$$

The main result can then be stated as follows (Cohen, Dahmen and DeVore, 2002c).

Theorem 8. Given any v and \mathcal{T}_ϵ defined by (157), we have the error estimate

$$\|\mathbf{F}(v) - \mathbf{F}(v)|_{\mathcal{T}_\epsilon}\|_{L_2} \lesssim \epsilon \quad (158)$$

Moreover, if $v \in \mathcal{A}_{\text{tree}}^s(L_2)$ with $0 < s < [(2\gamma - d)/2d]$, then we have the estimate

$$\#(\mathcal{T}_\epsilon) \lesssim \|v\|_{\mathcal{A}_{\text{tree}}^s(L_2)}^{1/\alpha} \epsilon^{-1/\alpha} + \#(\mathcal{T}_0) \quad (159)$$

We therefore have $\mathbf{F}(v) \in \mathcal{A}_{\text{tree}}^s(L_2)$ and

$$\|\mathbf{F}(v)\|_{\mathcal{A}_{\text{tree}}^s(L_2)} \lesssim 1 + \|v\|_{\mathcal{A}_{\text{tree}}^s(L_2)} \quad (160)$$

The constants in these above inequalities depend only on $\|v\|$, the space dimension d , and the parameter s .

This result provides the basis for the following evaluation scheme.

EVAL $\{\epsilon, \mathbf{F}, v\} \rightarrow w(\epsilon)$ PRODUCE FOR ANY FINITELY SUPPORTED VECTOR v A FINITELY SUPPORTED VECTOR $w(\epsilon)$ SUCH THAT $\|w(\epsilon) - \mathbf{F}(v)\|_{L_2} \leq \epsilon$ USING THE FOLLOWING STEPS:

(1) INVOKE THE ALGORITHM IN (BINEV AND DEVORE, 2002) TO COMPUTE THE TREES

$$\mathcal{T}_j := \mathcal{T}\left(\frac{2^j \epsilon}{C_0(j+1)}, v\right) \quad (161)$$

WHERE $C_0 = C_0(\|v\|)$ IS THE CONSTANT INVOLVED IN (158), FOR $j = 0, \dots, J$, AND STOP FOR THE SMALLEST J SUCH THAT \mathcal{T}_j IS EMPTY (WE ALWAYS HAVE $J \lesssim \log_2(\|v\|/\epsilon)$).

(2) DERIVE THE EXPANDED TREES $\tilde{\mathcal{T}}_j$, THE LAYERS Δ_j AND THE OUTCOME TREE \mathcal{T}_ϵ ACCORDING TO (157).

(3) COMPUTE $\mathbf{F}(v)|_{\mathcal{T}_\epsilon}$ (APPROXIMATELY WITHIN ACCURACY ϵ).

Clearly any finitely supported v belongs to $\mathcal{A}_{\text{tree}}^s(L_2)$ for every $s > 0$. Moreover, the trees \mathcal{T}_j will be empty

for $j \geq J$ and some $J \in \mathbb{N}$. Thus the scheme terminates after finitely many steps. We postpone some comments on step (3) to Section 6.4. The following theorem summarizes the properties of Algorithm EVAL.

Theorem 9. Given the inputs $\epsilon > 0$, a nonlinear function \mathbf{F} (such that \mathbf{F} satisfies assumptions of the type mentioned before), and a finitely supported vector v , then the output tree \mathcal{T}_ϵ has the following properties: One has $\|\mathbf{F}(v) - \mathbf{F}(v)|_{\mathcal{T}_\epsilon}\|_{L_2} \leq \epsilon$. Furthermore, for any $0 < s < [(2\gamma - d)/2d]$ (see Theorem 8), one has

$$\#(\mathcal{T}_\epsilon) \leq C \|v\|_{\mathcal{A}_{\text{tree}}^s(L_2)}^{1/\alpha} \epsilon^{-1/\alpha} + \#(\mathcal{T}_0) =: N_\epsilon \quad (162)$$

with C a constant depending only on the constants appearing in Theorem 8. Moreover, the number of computations needed to find \mathcal{T}_ϵ is bounded by $C(N_\epsilon + \#(\mathcal{T}(v)))$, where N_ϵ is the right hand side of (162) and $\mathcal{T}(v)$ is the smallest tree containing $\text{supp } v$.

Finally, we need a particular coarsening strategy that respects tree structures.

COARSE $[\eta, w] \rightarrow \tilde{w}_\eta$ DETERMINES FOR A FIXED CONSTANT $C^* \geq 1$, ANY FINITELY SUPPORTED INPUT w , AND ANY TOLERANCE $\eta > 0$ AN (η, C^*) -NEAR BEST TREE $\mathcal{T}(\eta, w)$ AND SETS $\tilde{w}_\eta := w|_{\mathcal{T}(\eta, w)}$.

The realization of COARSE is based on the results in Binev and DeVore (2002), which ensure linear complexity. This version of COARSE can also be used in the linear case. As such, it can be used to show that Theorem 7 remains valid for compressible matrices when the spaces $\mathcal{A}^s(L_2)$ are replaced by $\mathcal{A}_{\text{tree}}^s(L_2)$ (see Cohen, Dahmen and DeVore, 2002b).

The above results allow one to show that the scheme RES, in all the above examples satisfies the following:

Whenever the exact solution u of (90) belongs to $\mathcal{A}_{\text{tree}}^s(\mathcal{H})$ for some $s < s^*$, then one has for any finitely supported input v and any tolerance $\eta > 0$ that the output $r_\eta := \text{RES}(\eta, C, \mathbf{F}, v)$ satisfies

$$\begin{aligned} \# \text{supp } r_\eta &\leq C \eta^{-1/\alpha} \left(\|v\|_{\mathcal{A}_{\text{tree}}^s(L_2)}^{1/\alpha} + \|u\|_{\mathcal{A}_{\text{tree}}^s(L_2)}^{1/\alpha} + 1 \right) \\ \|r_\eta\|_{\mathcal{A}_{\text{tree}}^s(L_2)} &\leq C \left(\|v\|_{\mathcal{A}_{\text{tree}}^s(L_2)} + \|u\|_{\mathcal{A}_{\text{tree}}^s(L_2)} + 1 \right) \end{aligned} \quad (163)$$

where (in addition to the dependence given in the previous theorems) C depends only on s when $s \rightarrow s^*$. Moreover, the number of operations needed to compute w_η stays proportional to $\# \text{supp } r_\eta$.

One can show that the number of perturbed updates in step m of SOLVE executed before branching off into a coarsening step m), remains uniformly bounded independent of

the data and of the target accuracy ϵ . Therefore, the s^* -sparsity of the routine RES ensures that the $\mathcal{A}_{\text{tree}}^s(L_2)$ -norms of the approximations v^ϵ remain bounded in each update block m). The coarsening step is applied exactly in order to prevent the constants in these estimates from building up over several subsequent update blocks m). This is the consequence of the following Coarsening Lemma (Cohen, Dahmen and DeVore, 2002b).

Proposition 2. If $v \in \mathcal{A}_{\text{tree}}^s(L_2)$ and $\|v - w\|_{L_2} \leq \eta$ with $\# \text{supp } w < \infty$. Then $\tilde{w}_\eta := \text{COARSE}[2C^*\eta, w]$ satisfies

$$\# \text{supp } \tilde{w}_\eta \lesssim \|v\|_{\mathcal{A}_{\text{tree}}^s(L_2)}^{1/\alpha} \eta^{-1/\alpha}, \quad \|v - \tilde{w}_\eta\|_{L_2} \leq (1 + C^*)\eta$$

and

$$\|\tilde{w}_\eta\|_{\mathcal{A}_{\text{tree}}^s(L_2)} \lesssim \|v\|_{\mathcal{A}_{\text{tree}}^s(L_2)}$$

where C^* is the constant from the near-best tree construction scheme in Binev and DeVore (2002).

One can then use the above building blocks to show that SOLVE is optimal in the following sense (Cohen, Dahmen and DeVore, 2002b).

Theorem 10. If the exact solution $u = \sum_{\lambda \in \mathcal{J}} u_\lambda \psi_\lambda$ belongs to $\mathcal{A}_{\text{tree}}^s(\mathcal{H})$, for any $s < s^*(F, \Psi)$, then, for any target accuracy $\epsilon > 0$, the approximations $u(\epsilon)$ produced by SOLVE satisfy

$$\left\| u - \sum_{\lambda} \tilde{u}(\epsilon)_\lambda \psi_\lambda \right\|_{\mathcal{H}} \leq C_\Psi \epsilon$$

and

$$\# \text{supp } \tilde{u}(\epsilon), \text{ comp. work} \lesssim \epsilon^{-1/\alpha}$$

The above results cover the examples from Section 5.2, in particular, the mixed formulations. Note that there is no restriction on the choice of wavelet bases for the different solution components such as velocity and pressure in the Stokes problem. In contrast, in the classical approach, a compatible choice verifying the LBB condition is essential. In the adaptive context such constraints become void. In addition to these qualitative asymptotic results, the experiments in Dahlike, Dahmen and Urban (2002) and Dahmen, Urban and Vorloeper (2002) show that also quantitatively the performance of the adaptive scheme stays the same even when wavelet bases for velocity and pressure are used that, in connection with an a priori choice of finite dimensional trial spaces, would violate the LBB condition.

6.3 The Newton scheme

The scheme *Solve* is based on an ideal iteration of order one. The choice $C_n := (DF(u^n))^{-1}$ offers in principle a better convergence behavior of the outer iteration. In fact, for problems of the type (105) one can show that the Newton iteration (121) converges quadratically for a sufficiently good initial guess u^0 . On the other hand, it is not clear what the cost of each linear subproblem

$$DF(u^n)w^n = -(F(u^n) - f) \quad (164)$$

will amount to. A detailed analysis is given in Cohen, Dahmen and DeVore (2002b) where it is shown that the perturbed Newton iteration still retains a quadratic convergence while preserving an overall asymptotically optimal complexity in the sense of Theorem 10. It is perhaps worth stressing the following two points. The well-posedness (109) ensures that the problem (164) is well-conditioned, which suggests employing *Solve* for its approximate solution. Nevertheless, this raises two questions. First, the scheme *APPLY* needed to realize the residual approximation in *Solve* would require assembling in some sense the matrix $DF(u^n)$ in each update step with sufficient accuracy, which could be prohibitively expensive. However, the result of the application of $DF(u^n)$ to a vector w can be interpreted as the array of dual wavelet coefficients of a nonlinear composition with two wavelet expansions, since

$$DF(u^n)w = ((\psi_\lambda, DF(u^n)w))_{\lambda \in J} = ((\psi_\lambda, Q(u^n, w)))_{\lambda \in J} =: Q(u^n, w)$$

The approximation of $Q(u^n, w)$ can be realized again with the aid of the scheme *Eval* without assembling the Jacobian, which in this sense leads to a *matrix free* Newton scheme. In fact, in Cohen, Dahmen and DeVore (2002b) *Eval* is described for functions of several components. Further remarks on the related computational issues will be given in Section 6.4.

Secondly, the complexity analysis is not completely straightforward because one cannot directly infer from the sparseness of u to the sparseness of the Newton systems (164). However, one can show that these systems may be viewed as perturbations of another system whose solution is sparse whenever u is sparse. This, however, limits the target accuracy for (164). Nevertheless, one can show that this still suffices to ensure second order convergence of the outer iteration (121) (see Cohen, Dahmen and DeVore, 2002b).

6.4 Computational issues

The above complexity analysis works under the assumption that in the linear case the entries of A are computable at unit cost in order to invoke *APPLY*. Likewise in the nonlinear case, the entries of $F(v)$ in the predicted sets T_ϵ have to be computed. Under fairly general assumptions, both tasks can be handled by the following strategy. By definition, one has

$$\mathcal{F}(v) = \sum_{\lambda \in J} (\psi_\lambda, \mathcal{F}(v)) \tilde{\psi}_\lambda = \sum_{\lambda \in J} (F(v))_\lambda \tilde{\psi}_\lambda$$

where $\tilde{\psi}$ is the dual basis for Ψ and hence a Riesz basis for \mathcal{H}' . The idea is now to use an efficient *recovery scheme*, as described in Dahmen, Schneider and Xu (2000) and Barinka *et al.* (2003), that produces for a given target accuracy ϵ an approximation $g = \sum_{\lambda \in \Lambda_\epsilon} g_\lambda \tilde{\psi}_\lambda \in \mathcal{H}'$ to $\mathcal{F}(v)$ such that $\|\mathcal{F}(v) - g\|_{\mathcal{H}'} \leq \epsilon$ at a computational cost that stays proportional to the size of the prediction set T_ϵ from *Eval*. The norm equivalence now guarantees that the corresponding coefficient arrays exhibit essentially the same accuracy $\|F(v) - g\|_{\ell_2} \leq c_\Psi^{-1} \epsilon$. The important point is that individual coefficients are never computed but, solely based on the knowledge of the prediction set T_ϵ , quadrature is always used to generate on the function side an approximation to the whole object $\mathcal{F}(v)$ by local quasi-interpolant techniques to be able to keep the computational work proportional to the number of current degrees of freedom. This strategy justifies the above assumptions in a wide range of cases (see Barinka, 2003; Dahmen, Schneider and Xu, 2000; Barinka *et al.*, 2003 for details).

6.5 Concluding remarks

The primary goal of this section was to bring out the essential mechanisms and the potential of wavelet-based multiscale techniques and to understand under which circumstances optimal complexity can be achieved. A crucial role was played by the mapping properties of the underlying variational problem in conjunction with the availability of a wavelet Riesz basis for the corresponding energy space. This also revealed where the principal difficulties may arise. Depending on the nature of \mathcal{H} , or when dealing with complex geometries the construction of a good basis, Ψ may be very difficult or even impossible. Poor constants in (30) would spoil the quantitative performance of *Solve* significantly. Likewise poor constants in (92) would have the same effect. In fact, these constants may be parameter dependent and further work in this direction is under progress. But at least, the analysis reveals what one should be looking for in each concrete case which might certainly

require much more additional work. More information on first computational studies for elliptic and indefinite problems can be found in Barinka *et al.* (2001), Dahlke, Dahmen and Urban (2002), and Dahmen, Urban and Vorloeper (2002).

The quantitative improvement of evaluation schemes like *Eval* in conjunction with the strategies in Barinka (2003), Dahmen, Schneider and Xu (2000), and Barinka *et al.* (2003) certainly plays an important role. But already the pure prediction result in Theorem 8 at least gives rigorous bounds on the effect of certain nonlinearities concerning the interaction of fine and coarse scales – a problem that is at the heart of many multiscale phenomena in technology and science.

Finally, the difficulties of finding stable pairs of trial functions in many mixed formulations may help one to appreciate the principal merits of techniques that inherit the stability properties of the original problem. In this context, the above multiscale techniques incorporate a natural stabilization effect.

ACKNOWLEDGMENT

We are very indebted to Frank Brankamp, Helmut Harbrecht, Siegfried Müller and Reinhold Schneider for providing us with illustrations and numerical results concerning the material discussed in this chapter.

NOTES

- [1] Throughout this chapter, we sometimes write $A \lesssim B$ to indicate the existence of a constant c such that $A \leq cB$ independent of any parameters on which A and B may depend. Moreover $A \sim B$ means that $A \lesssim B$ and $B \lesssim A$.
- [2] There is at least one strategy for maintaining the Euclidean structure by employing fictitious domain techniques; appending, for instance, essential boundary conditions by Lagrange multipliers (Dahmen and Kunoth, 2001; Kunoth, 1995).

REFERENCES

- Adams RA. *Sobolev Spaces*. Academic Press; New York, 1978.
- Alpert B. A class of bases in for sparse representation of integral operators. *SIAM J. Math. Anal.* 1993; 24:246–262.
- Arandiga F, Donat R and Harten A. Multiresolution based on weighted averages of the hat function I: linear reconstruction techniques. *SIAM J. Numer. Anal.* 1998; 36:160–203.

- Arandiga F, Donat R and Harten A. Multiresolution based on weighted averages of the hat function II: nonlinear reconstruction techniques. *SIAM Sci. Comput.* 1999; 20:1053–1093.
- Ballmann J, Brankamp F and Müller S. Development of a flow solver employing local adaptation based on multiscale analysis based on B-spline grids, in: *Proceedings of the 8th Annual Conference of the CFD Society of Canada*, Montreal, 11–13 June 2000.
- Brankamp F, Gottschlich-Müller B, Hesse M, Lamby Ph, Müller S, Ballmann J, Brakhage K-H and Dahmen W. H-adaptive multiscale schemes for the compressible Navier–Stokes equations – polyhedral discretization, data compression and mesh generation, notes on numerical fluid mechanics. In *Flow Modulation and Fluid-Structure-Interaction at Airplane Wings*, Ballmann J (ed.). Springer-Verlag: 2003; 125–204.
- Barinka A. *Fast Evaluation Tools for Adaptive Wavelet Schemes*. PhD thesis, RWTH, Aachen, 2004.
- Barinka A, Dahmen W and Schneider R. An Algorithm for the Evaluation of Nonlinear Functionals of Wavelet Expansions, 2004.
- Barinka A, Dahlke S and Dahmen W. Adaptive Application of Operators in Standard Representation, *Adv. Comp. Math. Preprint*, RWTH-Aachen, Oct. 2003.
- Barinka A, Bartsch T, Charton P, Cohen A, Dahlke S, Dahmen W and Urban K. Adaptive wavelet schemes for elliptic problems – implementation and numerical experiments. *SIAM J. Sci. Comput.* 2001; 23(3):910–939.
- Berger M and Olinger J. Adaptive mesh refinement for hyperbolic partial differential equations. *J. Comput. Phys.* 1984; 53:484–512.
- Bergh J and Löfström J. *Interpolation Spaces, An Introduction*. Springer: Berlin, 1976.
- Bertoluzza S. A-posteriori error estimates for wavelet Galerkin methods. *Appl. Math. Lett.* 1995; 8:1–6.
- Bertoluzza S. An adaptive collocation method based on interpolating wavelets. In *Multiscale Wavelet Methods for PDE's*, Dahmen W, Kurdila AJ and Oswald P (eds). Academic Press: New York, 1997; 109–135.
- Beylkin G, Coifman RR and Rokhlin V. Fast wavelet transforms and numerical algorithms I. *Commun. Pure Appl. Math.* 1991; 44:141–183.
- Binev P and DeVore R. Fast Computation in Adaptive Tree Approximation. University of South Carolina. *Numerische Mathematik*, 2004; 97:193–217.
- Bramble JH, Lazarov RD and Pasciak JE. A least-squares approach based on a discrete minus one inner product for first order systems. *Math. Comp.* 1997; 66:935–955.
- Brezzi F and Fortin M. *Mixed and Hybrid Finite Element Methods*. Springer: Berlin, Heidelberg, New York, 1991.
- Canuto A, Tabacco A and Urban K. The wavelet element method, part I: construction and analysis. *Appl. Comput. Harm. Anal.* 1999; 6:1–52.
- Canuto A, Tabacco A and Urban K. The wavelet element method, part II: realization and additional features. *Appl. Comput. Harm. Anal.* 2000; 8:123–165.
- Carnicer JM, Dahmen W and Peña JM. Local decomposition of refinable spaces. *Appl. Comput. Harm. Anal.* 1996; 3:127–153.

- Cohen A. *Wavelet Methods in Numerical Analysis, in the Handbook of Numerical Analysis*, vol. VII, Ciarlet P-G et Lions J-L (eds). Elsevier: Amsterdam, 2000.
- Cohen A. *Numerical Analysis of Wavelet Methods, Studies in Mathematics and its Applications*, vol. 32. Elsevier: Amsterdam, 2003.
- Cohen A, Daubechies I and Feuvreau J-C. Biorthogonal bases of compactly supported wavelets. *Commun. Pure Appl. Math.* 1992; 45:485–560.
- Cohen A, Dahmen W and DeVore R. Adaptive wavelet methods for elliptic operator equations – convergence rates. *Math. Comp.* 2001; 70:27–75.
- Cohen A, Dahmen W and DeVore R. Adaptive wavelet methods II – beyond the elliptic case. *Found. Comput. Math.* 2002; 2(3):203–245.
- Cohen A, Dahmen W and DeVore R. Adaptive Wavelet Schemes for Nonlinear Variational Problems. *SIAM J. Numer. Anal.* 2003; 5(4):1783–1823.
- Cohen A, Dahmen W and DeVore R. Sparse Evaluation of Compositions of Functions Using multiscale Expansions. *SIAM J. Math. Anal.* 2003; 35:279–303.
- Cohen A, Kaber S-M, Müller S and Pöschl M. Fully adaptive multiscale methods for conservation laws. *Math. Comp.* 2002; 72:183–225.
- Cohen A and Masson R. Wavelet adaptive methods for second order elliptic problems, boundary conditions and domain decomposition. *Numer. Math.* 1997; 8:21–47.
- Costabel M and Stephan EP. Coupling of finite and boundary element methods for an elastoplastic interface problem. *SIAM J. Numer. Anal.* 1990; 27:1212–1226.
- Dahike S and DeVore R. Besov regularity for elliptic boundary value problems. *Commun. Part. Differ. Equations* 1997; 22:1–16.
- Dahike S, Dahmen W and Urban K. Adaptive wavelet methods for saddle point problems – convergence rates. *SIAM J. Numer. Anal.* 2002; 40(4):1230–1262.
- Dahike S, Dahmen W, Hochmuth R and Schneider R. Stable multiscale bases and local error estimation for elliptic problems. *Appl. Numer. Math.* 1997; 8:21–47.
- Dahike S. Besov regularity for elliptic boundary value problems on polygonal domains. *Appl. Math. Lett.* 1999; 12:31–36.
- Dahmen W. Some remarks on multiscale transformations, stability and biorthogonality. In *Wavelets, Images and Surface Fitting*, Laurent PJ, Le Méhauté A and Schumaker LL (eds), AK Peters: Wellesley, 1994; 157–188.
- Dahmen W. Stability of multiscale transformations. *J. Fourier Anal. Appl.* 1996; 2:341–361.
- Dahmen W. Wavelet and multiscale methods for operator equations. *Acta Numer.*, 1997; 6:55–228.
- Dahmen W. Wavelet methods for PDEs – Some recent developments. *J. Comput. Appl. Math.* 2001; 128:133–185.
- Dahmen W. *Multiscale and Wavelet Methods for Operator Equations*, C.I.M.E. Lecture Notes. Springer-Verlag: Heidelberg 1825: 2003.
- Dahmen W, Gottschlich-Müller B and Müller S. Multiresolution schemes for conservation laws. *Numer. Math.* 2001; 88:399–443.
- Dahmen W and Kunoth A. Multilevel preconditioning. *Numer. Math.* 1992; 63:315–344.
- Dahmen W and Kunoth A. Appending boundary conditions by Lagrange multipliers: analysis of the LBB condition. *Numer. Math.* 2001; 88:9–42.
- Dahmen W and Kunoth A. *Adaptive Wavelet Methods for Linear-Quadratic Elliptic Control Problems: Convergence Rates*. IGPM Report # 224, RWTH Aachen, December 2002.
- Dahmen W and Schneider R. Composite wavelet bases for operator equations. *Math. Comp.* 1999; 68:1533–1567.
- Dahmen W and Schneider R. Wavelets on manifolds I: construction and domain decomposition. *SIAM J. Math. Anal.* 1999; 31:184–230.
- Dahmen W and Stevenson R. Element-by-element construction of wavelets – stability and moment conditions. *SIAM J. Numer. Anal.* 1999; 37:319–325.
- Dahmen W, Harbrecht H and Schneider R. *Compression Techniques for Boundary Integral Equations – Optimal Complexity Estimates*. IGPM Report # 218, RWTH Aachen, June 2002.
- Dahmen W, Kunoth A and Schneider R. Wavelet least squares methods for boundary value problems. *SIAM J. Numer. Anal.* 2002; 39(6):1985–2013.
- Dahmen W, Kunoth A and Urban K. Biorthogonal spline-wavelets on the interval – stability and moment conditions. *Appl. Comput. Harm. Anal.* 1999; 6:132–196.
- Dahmen W, Proßdorf S and Schneider R. Multiscale methods for pseudo-differential equations on smooth manifolds. In *Proceedings of the International Conference on Wavelets: Theory, Algorithms, and Applications*, Chui CK, Montefusco L and Puccio L (eds), Academic Press, 1994; 385–424.
- Dahmen W, Schneider R and Xu Y. Nonlinear functions of wavelet expansions – adaptive reconstruction and fast evaluation. *Numer. Math.* 2000; 86:49–101.
- Dahmen W, Urban K and Vorloper J. Adaptive wavelet methods – basic concepts and applications. In *Wavelet Analysis – Twenty Years Developments*, Ding-Xuan Zhou (ed.), World Scientific: New Jersey, London, Singapore, Hongkong, 2002; 39–80.
- Daubechies I. *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Math. 61, SIAM, Philadelphia, 1992.
- DeVore R. Nonlinear approximation. *Acta Numer.*, 1998; 7: 51–150.
- DeVore R and Lorentz GG. *Constructive Approximation*, Grundlehren vol. 303. Springer-Verlag: Berlin, 1993.
- DeVore R and Popov V. Interpolation of Besov spaces. *Trans. Am. Math. Soc.* 1988; 305:397–414.
- DeVore R, Jawerth B and Popov V. Compression of wavelet decompositions. *Am. J. Math.* 1992; 114:737–785.
- DeVore R and Sharpley R. Maximal functions measuring smoothness. *Memoirs Am. Math. Soc.* 1984; 47:1–115.
- Girault V and Raviart P-A. *Finite Element Methods for Navier-Stokes Equations*. Springer: Berlin Heidelberg New York, 1986.
- Greengard L and Rokhlin V. A fast algorithm for particle simulation. *J. Comput. Phys.* 1987; 73:325–348.
- Hackbusch W. A sparse matrix arithmetic based on \mathcal{H} -matrices. Part I: Introduction to \mathcal{H} -matrices. *Computing* 1999; 64: 89–108.
- Hackbusch W and Nowak ZP. On the fast matrix multiplication in the boundary element method by panel clustering. *Numer. Math.* 1989; 54:463–491.
- Harbrecht H. *Wavelet Galerkin Schemes for the Boundary Element Method in Three Dimensions*. PhD thesis, Technische Universität Chemnitz, Chemnitz 2001.
- Harbrecht H. private communication.
- Harten A. Discrete multiresolution and generalized wavelets. *J. Appl. Numer. Math.* 1993; 12:153–193.
- Harten A. Multiresolution algorithms for the numerical solution of hyperbolic conservation laws. *Commun. Pure Appl. Math.* 1995; 48:1305–1342.
- Harten A. Multiresolution representation of data II: generalized framework. *SIAM J. Num. Anal.* 1996; 33:1205–1256.
- Lage C. Concept oriented design of numerical software. *Boundary Elements: Implementation and Analysis of Advanced Algorithms; Proceedings of the 12th GAMM-Seminar*, Kiel, 19–21 January, 1996; Vieweg Notes Numer. Fluid Mech. 1996; 54:159–170.
- Lage C and Schwab C. Wavelet Galerkin algorithms for boundary integral equations. *SIAM J. Sci. Statist. Comput.* 1998; 20:2195–2222.
- Kress R. *Linear Integral Equations*. Springer-Verlag: Berlin, Heidelberg, 1989.
- Kunoth A. Multilevel preconditioning – appending boundary conditions by Lagrange multipliers. *Adv. Comput. Math.* 1995; 4:145–170.
- Kunoth A. Wavelet methods – elliptic boundary value problems and control problems. In *Advances in Numerical Mathematics*, Bock HG, Hackbusch W, Luskin M and Hackbusch W (eds), Teubner: Stuttgart-Leipzig-Wiesbaden, 2001.
- Maday Y, Perrier V and Ravel JC. Adaptativité dynamique sur bases d'ondelettes pour l'approximation d'équations aux dérivées partielles. *C.R. Acad. Sci. Paris, Série* 1991; I: 405–410.
- Müller S. *Adaptive Multiscale Schemes for Conservation Laws, Lecture Notes in Computational Science and Engineering*, vol. 27. Springer-Verlag: Berlin, Heidelberg, 2003.
- von Petersdorff T and Schwab C. Wavelet approximation for first kind integral equations on polygons. *Numer. Math.* 1996; 74:479–519.
- von Petersdorff T and Schwab C. Fully discrete multiscale Galerkin BEM. In *Multiscale Wavelet Methods for PDEs*, Dahmen W, Kurdila A and Oswald P (eds), Academic Press: San Diego, 1997; 287–346.
- von Petersdorff T, Schneider R and Schwab C. Multiwavelets for second kind integral equations. *SIAM J. Numer. Anal.* 1997; 34:2212–2227.
- Schmidlin G, Lage C and Schwab C. *Rapid Solution of First Kind Boundary Integral Equations in \mathbb{R}^3* . Research Report No. 2002–07, Seminar für Angewandte Mathematik, ETH Zürich.
- Schneider R. *Multiskalen- und Wavelet-Matrixkompression: Analysebasierte Methoden zur Lösung großer vollbesetzter Gleichungssysteme*. B.G. Teubner: Stuttgart, 1998.
- Stevenson R. *Locally Supported, Piecewise Polynomial Biorthogonal Wavelets on Non-Uniform Meshes*. Technical Report 1157, University of Utrecht: Utrecht, 2000, to appear in Constructive Approximation.
- Stevenson R. *On the Compressibility of Operators in Wavelet Coordinates*. Technical Report 1249, University of Utrecht: Utrecht, 2002, to appear in SIAM J. Math. Anal.
- Sweldens W. The lifting scheme: a custom-design construction of biorthogonal wavelets. *Appl. Comput. Harm. Anal.* 1996; 3:186–200.
- Sweldens W. The lifting scheme: a construction of second generation wavelets. *SIAM J. Math. Anal.* 1998; 29:511–546.
- Triebel H. *Interpolation Theory, Function Spaces, and Differential Operators*. North Holland: Amsterdam, 1978.
- Vassilevski PS and Wang J. Stabilizing the hierarchical basis by approximate wavelets. I: Theory. *Numer. Lin. Alg. Appl.* 1997; 4:103–126.
- Yserentant H. On the multilevel splitting of finite element spaces. *Numer. Math.* 1986; 49:379–412.

Chapter 8

Plates and Shells: Asymptotic Expansions and Hierarchic Models

Monique Dauge¹, Erwan Faou² and Zohar Yosibash³

¹IRMAR, Université de Rennes 1, Campus de Beaulieu, Rennes, France

²INRIA Rennes, Campus de Beaulieu, Rennes, France

³Department of Mechanical Engineering, Ben-Gurion University, Beer Sheva, Israel

| | |
|--|-----|
| 1 Introduction | 199 |
| 2 Multiscale Expansions for Plates | 202 |
| 3 Hierarchical Models for Plates | 207 |
| 4 Multiscale Expansions and Limiting Models for Shells | 211 |
| 5 Hierarchical Models for Shells | 218 |
| 6 Finite Element Methods in Thin Domains | 219 |
| Acknowledgments | 229 |
| Notes | 229 |
| References | 229 |
| Further Reading | 232 |

1 INTRODUCTION

1.1 Structures

Plates and shells are characterized by (i) their midsurface S , (ii) their thickness d . The plate or shell character is that d is *small* compared to the dimensions of S . In this respect, we qualify such structures as *thin domains*. In the case of plates, S is a domain of the plane, whereas in the case of shells, S is a surface embedded in the three-dimensional space. Of course, plates are shells with zero curvature.

Encyclopedia of Computational Mechanics, Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes. Volume 1: *Fundamentals*. © 2004 John Wiley & Sons, Ltd. ISBN: 0-470-84699-2.

Nevertheless, considering plates as a particular class of shells is not so obvious: They have always been treated separately, for the reason that plates are simpler. We think, and hopefully demonstrate in this chapter, that eventually, considering plates as shells sheds some light in the shell theory.

Other classes of thin domains do exist, such as rods, where two dimensions are small compared to the third one. We will not address them and quote, for example, (Nazarov, 1999; Irago and Viaño, 1999). Real engineering structures are often the union (or junction) of plates, rods, shells, and so on. See Ciarlet (1988, 1997) and also Kozlov, Maz'ya and Movchan (1999) and Agratov and Nazarov (2000). We restrict our analysis to an isolated plate or shell. We assume moreover that the midsurface S is smooth, orientable, and has a smooth boundary ∂S . The shell character includes the fact that the principal curvatures have the same order of magnitude as the dimensions of S . See Anicic and Léger (1999) for a situation where a region with strong curvature (like $1/d$) is considered. The opposite situation is when the curvatures have the order of d : We are then in the presence of shallow shells according to the terminology of Ciarlet and Paumier (1986).

1.2 Domains and coordinates

In connection with our references, it is easier for us to consider d as the *half-thickness* of the structure. We denote our plate or shell by Ω^d . We keep the reference to the half-thickness in the notation because we are going to perform

an asymptotic analysis for which we embed our structure in a whole family of structures $(\Omega^\varepsilon)_\varepsilon$, where the parameter ε tends to 0.

We denote the Cartesian coordinates of \mathbb{R}^3 by $x = (x_1, x_2, x_3)$, a tangential system of coordinates on S by $x_\tau = (x_\alpha)_{\alpha=1,2}$, a normal coordinate to S by x_3 , with the convention that the midsurface is parametrized by the equation $x_3 = 0$. In the case of plates (x_α) are Cartesian coordinates in \mathbb{R}^2 and the domain Ω^ε has the tensor product form

$$\Omega^\varepsilon = S \times (-d, d)$$

In the case of shells, $x_\tau = (x_\alpha)_{\alpha=1,2}$ denotes a local coordinate system on S , depending on the choice of a local chart in an atlas, and x_3 is the coordinate along a smooth unit normal field n to S in \mathbb{R}^3 . Such a normal coordinate system (also called S -coordinate system) (x_τ, x_3) yields a smooth diffeomorphism between Ω^ε and $S \times (-d, d)$. The lateral boundary Γ^ε of Ω^ε is characterized by $x_\tau \in \partial S$ and $x_3 \in (-d, d)$ in coordinates (x_τ, x_3) .

1.3 Displacement, strain, stress, and elastic energy

The displacement of the structure (deformation from the stress-free configuration) is denoted by u , its Cartesian coordinates by (u_1, u_2, u_3) , and its surface and transverse parts by $u_\tau = (u_\alpha)$ and u_3 respectively. The transverse part u_3 is always an intrinsic function and the surface part u_τ defines a two-dimensional 1-form field on S , depending on x_3 . The components (u_α) of u_τ depend on the choice of the local coordinate system x_τ .

We choose to work in the framework of small deformations (see Ciarlet (1997, 2000) for more general nonlinear models e.g. the von Kármán model). Thus, we use the strain tensor (linearized from the Green–St Venant strain tensor) $\varepsilon = (\varepsilon_{ij})$ given in Cartesian coordinates by

$$\varepsilon_{ij}(u) = \frac{1}{2} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)$$

Unless stated otherwise, we assume the simplest possible behavior for the material of our structure, that is, an isotropic material. Thus, the elasticity tensor $A = (A^{ijkl})$ takes the form

$$A^{ijkl} = \lambda \delta^{ij} \delta^{kl} + \mu (\delta^{ik} \delta^{jl} + \delta^{il} \delta^{jk})$$

with λ and μ the Lamé constants of the material and δ^{ij} the Kronecker symbol. We use Einstein's summation

convention, and sum over double indices if they appear as subscripts and superscripts (which is nothing but the contraction of tensors), for example, $\sigma^{ij} e_{ij} \equiv \sum_{i,j=1}^3 \sigma^{ij} e_{ij}$. The constitutive equation is given by Hooke's law $\sigma = A\varepsilon(u)$ linking the stress tensor σ to the strain tensor $\varepsilon(u)$. Thus

$$\begin{aligned} \sigma^{ii} &= \lambda(e_{11} + e_{22} + e_{33}) + 2\mu e_{ii}, \quad i = 1, 2, 3 \\ \sigma^{ij} &= 2\mu e_{ij} \quad \text{for } i \neq j \end{aligned} \quad (1)$$

The elastic bilinear form on a domain Ω is given by

$$a(u, u') = \int_\Omega \sigma(u) : \varepsilon(u') \, dx = \int_\Omega \sigma^{ij}(u) \varepsilon_{ij}(u') \, dx \quad (2)$$

and the elastic energy of a displacement u is $(1/2)a(u, u)$. The strain–energy norm of u is denoted by $\|u\|_{\mathcal{E}(\Omega)}$ and defined as $(\sum_{i,j} \int_\Omega |\varepsilon_{ij}(u)|^2 \, dx)^{1/2}$.

1.4 Families of problems

We will address two types of problems on our thin domain Ω^ε : (i) Find the displacement u solution to the equilibrium equation $\operatorname{div} \sigma(u) = f$ for a given load f , (ii) Find the (smallest) vibration eigen-modes (Λ, u) of the structure. For simplicity of exposition, we assume in general that the structure is clamped (this condition is also called ‘condition of place’) along its lateral boundary Γ^ε and will comment on other choices for lateral boundary conditions. On the remaining part of the boundary $\partial\Omega^\varepsilon \setminus \Gamma^\varepsilon$ (‘top’ and ‘bottom’) traction free condition is assumed.

In order to investigate the influence of the thickness on the solutions and the discretization methods, we consider our (fixed physical) problem in Ω^ε as part of a whole family of problems, depending on one parameter $\varepsilon \in (0, \varepsilon_0]$, the thickness. The definition of Ω^ε is obvious by the formulae given in Section 1.2 (in fact, if the curvatures of S are ‘small’, we may decide that Ω^ε fits better in a family of shallow shells, see Section 4.4 later). For problem (i), we choose the same right hand side f for all values of ε , which precisely means that we fix a smooth field f on Ω^0 and take $f^\varepsilon := f|_{\Omega^\varepsilon}$ for each ε .

Both problems (i) and (ii) can be set in variational form (principle of virtual work). Our three-dimensional variational space is the subspace $V(\Omega^\varepsilon)$ of the Sobolev space $H^1(\Omega^\varepsilon)^3$ characterized by the clamping condition $u|_{\Gamma^\varepsilon} = 0$, and the bilinear form a (2) on $\Omega = \Omega^0$, denoted by a^ε . The variational formulations are

Find $u^\varepsilon \in V(\Omega^\varepsilon)$ such that

$$a^\varepsilon(u^\varepsilon, u') = \int_{\Omega^\varepsilon} f^\varepsilon \cdot u' \, dx, \quad \forall u' \in V(\Omega^\varepsilon) \quad (3)$$

for the problem with external load, and

Find $u^\varepsilon \in V(\Omega^\varepsilon)$, $u^\varepsilon \neq 0$, and $\Lambda^\varepsilon \in \mathbb{R}$ such that

$$a^\varepsilon(u^\varepsilon, u') = \Lambda^\varepsilon \int_{\Omega^\varepsilon} u^\varepsilon \cdot u' \, dx, \quad \forall u' \in V(\Omega^\varepsilon) \quad (4)$$

for the eigen-mode problem. In engineering practice, one is interested in the natural frequencies, $\omega^\varepsilon = \frac{1}{2\pi} \sqrt{\Lambda^\varepsilon}$. Of course, when considering our structure Ω^ε , we are eventually only interested in $\varepsilon = d$. Taking the whole family $\varepsilon \in (0, \varepsilon_0]$ into account allows the investigation of the dependency with respect to the small parameter ε , in order to know if valid simplified models are available and how they can be discretized by finite elements.

1.5 Computational obstacles

Our aim is to study the possible discretizations for a reliable and efficient computation of the solutions u^ε of problem (3) or (4) in our thin structure Ω^ε . An option could be to consider Ω^ε as a three-dimensional body and use 3-D finite elements. In the standard version of finite elements (h -version), individual elements should not be stretched or distorted, which implies that all dimensions should be bounded by d . Even so, several layers of elements through the thickness may be necessary. Moreover the a priori error estimates may suffer from the behavior of the Korn inequality on Ω^ε (the factor appearing in the Korn inequality behaves like d^{-1} for plates and partially clamped shells; see Ciarlet, Lods and Miara (1996) and Dauge and Faou (2004)).

An ideal alternative would simply be to get rid of the thickness variable and compute the solution of an ‘equivalent’ problem on the midsurface S . This is the aim of the shell theory. Many investigations were undertaken around 1960–1970, and the main achievement is (still) the Koiter model, which is a multidegree 3×3 elliptic system on S of half-orders $(1, 1, 2)$ with a singular dependence in d . But, as written in Koiter and Simmonds (1973), ‘Shell theory attempts the impossible: to provide a two-dimensional representation of an intrinsically three-dimensional phenomenon’. Nevertheless, obtaining converging error estimates between the 3-D solution u^ε and a reconstructed 3-D displacement U^ε from the deformation pattern z^ε solution of the Koiter model seems possible.

However, due to its fourth order part, the Koiter model cannot be discretized by standard C^0 finite elements. The

Naghdi model, involving five unknowns on S , seems more suitable. Yet, endless difficulties arise in the form of various locking effects, due to the singularly perturbed character of the problem.

With the twofold aim of improving the precision of the models and their approximability by finite elements, the idea of hierarchical models becomes natural: Roughly, it consists of an Ansatz of polynomial behavior in the thickness variable, with bounds on the degrees of the three components of the 3-D displacement. The introduction of such models in variational form is due to Vogelius and Babuška (1981c) and Szabó and Sahrman (1988). Earlier beginnings in that direction can be found in Vekua (1955, 1965). The hierarchy (increasing the transverse degrees) of models obtained in that way can be discretized by the p -version of finite elements.

1.6 Plan of the chapter

In order to assess the validity of hierarchical models, we will compare them with asymptotic expansions of solutions u^ε when they are available: These expansions exhibit two or three different scales and boundary layer regions, which can or cannot be properly described by hierarchical models.

We first address plates because much more is known for plates than for general shells. In Section 2, we describe the two-scale expansion of the solutions of (3) and (4): This expansion contains (i) a regular part each term of which is polynomial in the thickness variable x_3 , (ii) a part mainly supported in a boundary layer around the lateral boundary Γ^ε . In Section 3, we introduce the hierarchical models as Galerkin projections on semidiscrete subspaces $V^q(\Omega^\varepsilon)$ of $V(\Omega^\varepsilon)$ defined by assuming a polynomial behavior of degree $q = (q_1, q_2, q_3)$ in x_3 . The model of degree $(1, 1, 0)$ is the Reissner–Mindlin model and needs the introduction of a reduced energy. The $(1, 1, 2)$ model is the lowest degree model to use the same elastic energy (2) as the 3-D model.

We address shells in Section 4 (asymptotic expansions and limiting models) and Section 5 (hierarchical models). After a short introduction of the metric and curvature tensors on the midsurface, we first describe the three-scale expansion of the solutions of (3) on clamped elliptic shells: Two of these scales can be captured by hierarchical models. We then present and comment on the famous classification of shells as flexural or membrane. We also mention two distinct notions of shallow shells. We emphasize the universal role played by the Koiter model for the structure Ω^ε , independently of any embedding of Ω^ε in a family $(\Omega^\varepsilon)_\varepsilon$.

The last section is devoted to the discretization of the 3-D problems and their 2-D hierarchical projections, by p -version finite elements. The 3-D thin elements (one layer of

elements through the thickness) constitute a bridge between 3-D and 2-D discretizations. We address the issue of locking effects (shear and membrane locking) and the issue of capturing boundary layer terms. Increasing the degree p of approximation polynomials and using anisotropic meshes is a way toward solving these problems. We end this chapter by presenting a series of eigen-frequency computations on a few different families of shells and draw some 'practical' conclusions.

2 MULTISCALE EXPANSIONS FOR PLATES

The question of an asymptotic expansion for solutions \mathbf{u}^ε of problems (3) or (4) posed in a family of plates is difficult: One may think it is natural to expand \mathbf{u}^ε either in polynomial functions in the thickness variable x_3 , or in an asymptotic series in powers ε^k with regular coefficients \mathbf{v}^k defined on the stretched plate $\Omega = S \times (-1, 1)$. In fact, for the class of loads considered here or for the eigen-mode problem, both those *Ansätze* are relevant, but they are unable to provide a correct description of the behavior of \mathbf{u}^ε in the vicinity of the lateral boundary Γ^ε , where there is a boundary layer of width $\sim \varepsilon$ (except in the particular situation of a rectangular midsurface with symmetry lateral boundary conditions (hard simple support or sliding edge); see Paumier, 1990). And, worse, in the absence of knowledge of the boundary layer behavior, the determination of the terms \mathbf{v}^k is impossible (except for \mathbf{v}^0).

The investigation of asymptotics as $\varepsilon \rightarrow 0$ was first performed by the construction of infinite formal expansions; see Friedrichs and Dressler (1961), Gol'denveizer (1962), and Gregory and Wan (1984). The principle of multiscale asymptotic expansion is applied to thin domains in Maz'ya, Nazarov and Plamenevskii (1991b). A two-term asymptotics is exhibited in Nazarov and Zorin (1989). The whole asymptotic expansion is constructed in Dauge and Gruais (1996, 1998a) and Dauge, Gruais and Rösle (1999/00).

The multiscale expansions that we propose differ from the matching method in Il'in (1992) where the solutions of singularly perturbed problems are fully described in rapid variables inside the boundary layer and slow variables outside the layer, both expansions being 'matched' in an intermediate region. Our approach is closer to that of Vishik and Lyusternik (1962) and Oleinik, Shamaev and Yosifian (1992) (see Chapter 2, Chapter 3 of Volume 2).

2.1 Coordinates and symmetries

The midsurface S is a smooth domain of the plane $\Pi \simeq \mathbb{R}^2$ (see Fig. 1) and for $\varepsilon \in (0, \varepsilon_0)$ $\Omega^\varepsilon = S \times (-\varepsilon, \varepsilon)$ is the

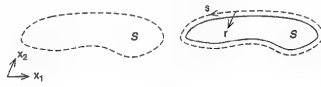


Figure 1. Cartesian and local coordinates on the midsurface.

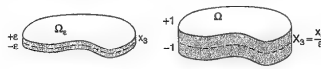


Figure 2. Thin plate and stretched plate.

generic member of the family of plates (see Fig. 2). The plates are symmetric with respect to the plane Π . Since they are assumed to be made of an isotropic material, problems (3) or (4) commute with the symmetry $\mathcal{G}: \mathbf{u} \mapsto (\mathbf{u}_T(\cdot, -x_3), -u_3(\cdot, -x_3))$. The eigenspaces of \mathcal{G} are *membrane* and *bending* displacements (also called stretching and flexural displacements), cf. Friedrichs and Dressler (1961):

$$\begin{aligned} \mathbf{u} \text{ membrane iff } & \mathbf{u}_T(\mathbf{x}_T, +x_3) = \mathbf{u}_T(\mathbf{x}_T, -x_3) \\ & \text{and } u_3(\mathbf{x}_T, +x_3) = -u_3(\mathbf{x}_T, -x_3) \\ \mathbf{u} \text{ bending iff } & \mathbf{u}_T(\mathbf{x}_T, +x_3) = -\mathbf{u}_T(\mathbf{x}_T, -x_3) \\ & \text{and } u_3(\mathbf{x}_T, +x_3) = u_3(\mathbf{x}_T, -x_3) \end{aligned} \quad (5)$$

Any general displacement \mathbf{u} is the sum $\mathbf{u}_m + \mathbf{u}_b$ of a membrane and a bending part (according to formulae $\mathbf{u}_m = (1/2)(\mathbf{u} + \mathcal{G}\mathbf{u})$ and $\mathbf{u}_b = (1/2)(\mathbf{u} - \mathcal{G}\mathbf{u})$). They are also denoted by \mathbf{u}^m and \mathbf{u}^b in the literature).

In addition to the coordinates \mathbf{x}_T in S , let r be the distance to ∂S in Π and s an arclength function on ∂S (see Fig. 1). In this way, (r, s) defines a smooth coordinate system in a midplane tubular neighborhood \mathcal{V} of ∂S . Let $\chi = \chi(r)$ be a smooth cut-off function with support in \mathcal{V} , equal to 1 in a smaller such neighborhood. It is used to subordinate boundary layer terms. The two following stretched (or rapid) variables appear in our expansions:

$$X_3 = \frac{x_3}{\varepsilon} \quad \text{and} \quad R = \frac{r}{\varepsilon}$$

The stretched thickness variable X_3 belongs to $(-1, 1)$ and is present in all parts of our asymptotics, whereas the presence of R characterizes boundary layer terms (see Figure 2).

2.2 Problem with external load

The solutions of the family of problems (3) have a two-scale asymptotic expansion in regular terms \mathbf{v}^k and boundary

layer terms \mathbf{w}^k , which we state as a theorem (Dauge, Gruais and Rösle, 1999/00; Dauge and Schwab, 2002). Note that in contrast with the most part of those references, we work here with *natural displacements* (i.e. unscaled), which is more realistic from the mechanical and computational point of view, and allows an easier comparison with shells.

Theorem 1. (Dauge, Gruais and Rösle, 1999/00) For the solutions of problem (3), $\varepsilon \in (0, \varepsilon_0]$, there exist regular terms $\mathbf{v}^k = \mathbf{v}^k(\mathbf{x}_T, X_3)$, $k \geq -2$, and boundary layer terms $\mathbf{w}^k = \mathbf{w}^k(R, s, X_3)$, $k \geq 0$, such that

$$\begin{aligned} \mathbf{u}^\varepsilon \simeq & \varepsilon^{-2} \mathbf{v}^{-2} + \varepsilon^{-1} \mathbf{v}^{-1} + \varepsilon^0 (\mathbf{v}^0 + \chi \mathbf{w}^0) \\ & + \varepsilon^1 (\mathbf{v}^1 + \chi \mathbf{w}^1) + \dots \end{aligned} \quad (6)$$

in the sense of asymptotic expansions: The following estimates hold

$$\left\| \mathbf{u}^\varepsilon - \sum_{k=-2}^K \varepsilon^k (\mathbf{v}^k + \chi \mathbf{w}^k) \right\|_{G(\mathbb{R}^3)} \leq C_K(\eta) \varepsilon^{K+1/2}, \quad K = 0, 1, \dots$$

where we have set $\mathbf{w}^{-2} = \mathbf{w}^{-1} = 0$ and the constant $C_K(\eta)$ is independent of $\varepsilon \in (0, \varepsilon_0]$.

2.2.1 Kirchhoff displacements and their deformation patterns

The first terms in the expansion of \mathbf{u}^ε are Kirchhoff displacements, that is, displacements of the form (with the surface gradient $\nabla_T = (\partial_1, \partial_2)$)

$$(\mathbf{x}_T, X_3) \mapsto \mathbf{v}(\mathbf{x}_T, X_3) = (\xi_T(\mathbf{x}_T) - x_3 \nabla_T \xi_3(\mathbf{x}_T), \xi_3(\mathbf{x}_T)) \quad (7)$$

Here, $\xi_T = (\xi_\alpha)$ is a surface displacement and ξ_3 is a function on S . We call the three-component field $\xi := (\xi_T, \xi_3)$ the *deformation pattern* of the KL displacement \mathbf{v} . Note that

$$\mathbf{v} \text{ bending iff } \xi = (0, \xi_3) \quad \text{and} \quad \mathbf{v} \text{ membrane iff } \xi = (\xi_T, 0)$$

In expansion (6) the first terms are Kirchhoff displacements. The next regular terms \mathbf{v}^k are also generated by deformation patterns ξ^k via higher degree formulae than in (7). We successively describe the \mathbf{v}^k , the ξ^k and, finally, the boundary layer terms \mathbf{w}^k .

2.2.2 The four first regular terms

For the regular terms \mathbf{v}^k , $k = -2, -1, 0, 1$, there exist bending deformation patterns $\xi^{-2} = (0, \xi_3^{-2})$, $\xi^{-1} = (0, \xi_3^{-1})$,

and full deformation patterns ξ^0, ξ^1 such that

$$\begin{aligned} \mathbf{v}^{-2} &= (0, \xi_3^{-2}) \\ \mathbf{v}^{-1} &= (-X_3 \nabla_T \xi_3^{-2}, \xi_3^{-1}) \\ \mathbf{v}^0 &= (\xi_T^0 - X_3 \nabla_T \xi_3^{-1}, \xi_3^0) + (0, P_0^2(X_3) \Delta_T \xi_3^{-2}) \\ \mathbf{v}^1 &= (\xi_T^1 - X_3 \nabla_T \xi_3^0, \xi_3^1) + (P_1^2(X_3) \nabla_T \Delta_T \xi_3^{-2}, \\ & \quad P_0^1(X_3) \operatorname{div} \xi_T^0 + P_2^2(X_3) \Delta_T \xi_3^{-1}) \end{aligned} \quad (8)$$

In the above formulae, $\nabla_T = (\partial_1, \partial_2)$ is the surface gradient on S , $\Delta_T = \partial_1^2 + \partial_2^2$ is the surface Laplacian and $\operatorname{div} \xi_T$ is the surface divergence (i.e. $\operatorname{div} \xi_T = \partial_1 \xi_1 + \partial_2 \xi_2$). The functions P_0^k and P_1^k are polynomials of degree k , whose coefficients depend on the Lamé constants according to

$$\begin{aligned} P_0^1(X_3) &= -\frac{\lambda}{\lambda + 2\mu} X_3, \\ P_0^2(X_3) &= \frac{\lambda}{2\lambda + 4\mu} \left(X_3^2 - \frac{1}{3} \right), \\ P_1^3(X_3) &= \frac{1}{6\lambda + 12\mu} ((3\lambda + 4\mu) X_3^3 - (11\lambda + 12\mu) X_3). \end{aligned} \quad (9)$$

Note that the first blocks in $\sum_{k \geq -2} \varepsilon^k \mathbf{v}^k$ yield Kirchhoff displacements, whereas the second blocks have zero mean values through the thickness for each $\mathbf{x}_T \in S$.

2.2.3 All regular terms with the help of formal series

We see from (8) that the formulae describing the successive \mathbf{v}^k are *partly self-similar* and, also, that each \mathbf{v}^k is enriched by a new term. That is why the whole regular term series $\sum_k \varepsilon^k \mathbf{v}^k$ can be efficiently described with the help of the formal series product.

A formal series is an infinite sequence $(a^0, a^1, \dots, a^k, \dots)$ of coefficients, which can be denoted in a symbolic way by $a[\varepsilon] = \sum_{k \geq 0} \varepsilon^k a^k$, and the product $a[\varepsilon]b[\varepsilon]$ of the two formal series $a[\varepsilon]$ and $b[\varepsilon]$ is the formal series $c[\varepsilon]$ with coefficients $c^k = \sum_{0 \leq j+k \leq k} a^j b^{k-j}$. In other words, the equation $c[\varepsilon] = a[\varepsilon]b[\varepsilon]$ is equivalent to the series of equation $c^k = \sum_{0 \leq j+k \leq k} a^j b^{k-j}$, $\forall k$.

With this formalism, we have the following identity, which extends formulae (8):

$$\mathbf{v}[\varepsilon] = \mathbf{V}[\varepsilon]\xi[\varepsilon] + \mathbf{Q}[\varepsilon]\mathbf{f}[\varepsilon] \quad (10)$$

(i) $\xi[\varepsilon]$ is the formal series of Kirchhoff deformation patterns $\sum_{k \geq -2} \varepsilon^k \xi^k$ starting with $k = -2$.

(ii) $\mathbf{V}[\varepsilon]$ has operator valued coefficients \mathbf{V}^k , $k \geq 0$, acting from $C^\infty(\bar{S})^3$ into $C^\infty(\bar{S})^3$.

$$\begin{aligned}
\mathbf{v}^0_\zeta &= (\zeta_\tau, \zeta_s) \\
\mathbf{v}^1_\zeta &= (-X_3 \nabla_\tau \zeta_s, P^1_\zeta(X_3) \operatorname{div} \zeta_\tau) \\
\mathbf{v}^2_\zeta &= (P^2_\zeta(X_3) \nabla_\tau \operatorname{div} \zeta_\tau, P^2_\zeta(X_3) \Delta_\tau \zeta_s) \\
&\vdots \\
\mathbf{v}^{2j}_\zeta &= (P^{2j}_\zeta(X_3) \nabla_\tau \Delta_\tau^{j-1} \operatorname{div} \zeta_\tau, P^{2j}_\zeta(X_3) \Delta_\tau^j \zeta_s) \\
\mathbf{v}^{2j+1}_\zeta &= (P^{2j+1}_\zeta(X_3) \nabla_\tau \Delta_\tau^j \zeta_s, P^{2j+1}_\zeta(X_3) \Delta_\tau^j \operatorname{div} \zeta_\tau)
\end{aligned} \quad (11)$$

with P^k_ζ and P^k_m polynomials of degree ℓ (the first ones are given in (9)).

(iii) $\mathbf{f}[\varepsilon]$ is the Taylor series of \mathbf{f} around the surface $X_3 = 0$:

$$\mathbf{f}[\varepsilon] = \sum_{k \geq 0} \varepsilon^k \mathbf{f}^k \text{ with } \mathbf{f}^k(\mathbf{x}_\tau, X_3) = \frac{X_3^k}{k!} \frac{\partial^k \mathbf{f}}{\partial X_3^k} \Big|_{X_3=0}(\mathbf{x}_\tau) \quad (12)$$

(iv) $\mathbf{Q}[\varepsilon]$ has operator valued coefficients \mathbf{Q}^k acting from $C^\infty(\bar{\Omega})^3$ into itself. It starts at $k=2$ (we can see now that the four first equations given by equality (10) are $\mathbf{v}^{-2} = \mathbf{v}^0 \zeta^{-2}$, $\mathbf{v}^{-1} = \mathbf{v}^0 \zeta^{-1} + \mathbf{v}^1 \zeta^{-2}$, $\mathbf{v}^0 = \mathbf{v}^0 \zeta^0 + \mathbf{v}^1 \zeta^{-1} + \mathbf{v}^2 \zeta^{-2}$, $\mathbf{v}^1 = \mathbf{v}^0 \zeta^1 + \mathbf{v}^1 \zeta^0 + \mathbf{v}^2 \zeta^{-1} + \mathbf{v}^3 \zeta^{-2}$, which gives back (8)).

$$\mathbf{Q}[\varepsilon] = \sum_{k \geq 2} \varepsilon^k \mathbf{Q}^k \quad (13)$$

Each \mathbf{Q}^k is made of compositions of partial derivatives in the surface variables \mathbf{x}_τ , with integral operators in the scaled transverse variable. Each of them acts in a particular way between semipolynomial spaces $E^q(\Omega)$, $q \geq 0$, in the scaled domain Ω : We define for any integer q , $q \geq 0$

$$\begin{aligned}
E^q(\Omega) &= \left\{ \mathbf{v} \in C^\infty(\Omega)^3, \exists \mathbf{x}^q \in C^\infty(\bar{\Omega})^3, \mathbf{v}(\mathbf{x}_\tau, X_3) \right. \\
&\quad \left. = \sum_{m=0}^q X_3^m \mathbf{x}^m(\mathbf{x}_\tau) \right\} \quad (14)
\end{aligned}$$

Note that by (12), \mathbf{f}^k belongs to $E^k(\Omega)$.

Besides, for any $k \geq 2$, \mathbf{Q}^k acts from $E^q(\Omega)$ into $E^{q+k}(\Omega)$. The first term of the series $\mathbf{Q}[\varepsilon]\mathbf{f}[\varepsilon]$ is $\mathbf{Q}^2\mathbf{f}^0$ and we have:

$$\mathbf{Q}^2\mathbf{f}^0(\mathbf{x}_\tau, X_3) = \left(0, \frac{1-3X_3^2}{6\lambda+12\mu} \mathbf{f}^0_\tau(\mathbf{x}_\tau) \right)$$

As a consequence of formula (10), combined with the structure of each term, we find

Lemma 1. (Dauge and Schwab, 2002) *With the definition (14) for the semipolynomial space $E^q(\Omega)$, for any $k \geq -2$ the regular term \mathbf{v}^k belongs to $E^{k+2}(\Omega)$.*

2.2.4 Deformation patterns

From formula (9) extended by (10) we obtain explicit expressions for the regular parts \mathbf{v}^k provided we know the deformation patterns ζ^k . The latter solves boundary value problems on the midsurface S . Our multiscale expansion approach gives back the well-known equations of plates (the Kirchhoff-Love model and the plane stress model) completed by a whole series of boundary value problems.

(i) The first bending generator ζ_3^{-2} solves the Kirchhoff-Love model

$$\begin{aligned}
L_b \zeta_3^{-2}(\mathbf{x}_\tau) &= \mathbf{f}^0_\tau(\mathbf{x}_\tau), \quad \mathbf{x}_\tau \in S \quad \text{with} \quad \zeta_3^{-2}|_{\partial S} = 0, \\
\partial_n \zeta_3^{-2}|_{\partial S} &= 0
\end{aligned} \quad (15)$$

where L_b is the fourth-order operator

$$L_b := \frac{4\mu}{3} \frac{\lambda + \mu}{\lambda + 2\mu} \Delta_\tau^2 = \frac{1}{3} (\tilde{\lambda} + 2\mu) \Delta_\tau^2 \quad (16)$$

and \mathbf{n} the unit interior normal to ∂S . Here $\tilde{\lambda}$ is the 'averaged' Lamé constant

$$\tilde{\lambda} = \frac{2\lambda\mu}{\lambda + 2\mu} \quad (17)$$

(iii) The second bending generator ζ_3^{-1} is the solution of a similar problem

$$\begin{aligned}
L_b \zeta_3^{-1}(\mathbf{x}_\tau) &= 0, \quad \mathbf{x}_\tau \in S \quad \text{with} \quad \zeta_3^{-1}|_{\partial S} = 0, \\
\partial_n \zeta_3^{-1}|_{\partial S} &= c_{\lambda,\mu}^0 \Delta_\tau \zeta_3^{-2}
\end{aligned} \quad (18)$$

where $c_{\lambda,\mu}^0$ is a positive constant depending on the Lamé coefficients.

(iii) The membrane part ζ_τ^0 of the third deformation pattern solves the plane stress model

$$L_m \zeta_\tau^0(\mathbf{x}_\tau) = \mathbf{f}^0_\tau(\mathbf{x}_\tau), \quad \mathbf{x}_\tau \in S \quad \text{and} \quad \zeta_\tau^0|_{\partial S} = 0 \quad (19)$$

where L_m is the second-order 2×2 system

$$\begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \mapsto \begin{pmatrix} (\tilde{\lambda} + 2\mu)\partial_{11} + \mu\partial_{22} & (\tilde{\lambda} + \mu)\partial_{12} \\ (\tilde{\lambda} + \mu)\partial_{12} & \mu\partial_{11} + (\tilde{\lambda} + 2\mu)\partial_{22} \end{pmatrix} \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \quad (20)$$

(iv) Here, again, the whole series of equations over the series of deformation patterns $\sum_{k \geq -2} \varepsilon^k \zeta^k$ can be

written in a global way using the formal series product, as reduced equations on the midsurface:

$$\mathbf{L}[\varepsilon]\zeta[\varepsilon] = \mathbf{R}[\varepsilon]\mathbf{f}[\varepsilon] \text{ in } S \quad \text{with} \quad \mathbf{d}[\varepsilon]\zeta[\varepsilon] = 0 \text{ on } \partial S \quad (21)$$

Here, $\mathbf{L}[\varepsilon] = \mathbf{L}^0 + \varepsilon^2 \mathbf{L}^2 + \varepsilon^4 \mathbf{L}^4 + \dots$, with

$$\begin{aligned}
\mathbf{L}^0 \zeta &= \begin{pmatrix} L_m & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \zeta_\tau \\ \zeta_s \end{pmatrix} \\
\mathbf{L}^2 \zeta &= \begin{pmatrix} L_m^2 & 0 \\ 0 & L_b \end{pmatrix} \begin{pmatrix} \zeta_\tau \\ \zeta_s \end{pmatrix}, \dots
\end{aligned} \quad (22)$$

where $L_m^2 \zeta_\tau$ has the form $c \nabla_\tau \Delta_\tau \operatorname{div} \zeta_\tau$. The series of operators $\mathbf{R}[\varepsilon]$ starts at $k=0$ and acts from $C^\infty(\bar{\Omega})^3$ into $C^\infty(\bar{S})^3$. Its first coefficient is the mean-value operator

$$\mathbf{f} \mapsto \mathbf{R}^0 \mathbf{f} \quad \text{with} \quad \mathbf{R}^0 \mathbf{f}(\mathbf{x}_\tau) = \frac{1}{2} \int_{-1}^1 \mathbf{f}(\mathbf{x}_\tau, X_3) dX_3 \quad (23)$$

Finally, the coefficients of the operator series $\mathbf{d}[\varepsilon]$ are trace operators acting on ζ . The first terms are

$$\begin{aligned}
\mathbf{d}^0 \zeta &= \begin{pmatrix} \zeta_\tau \cdot \mathbf{n} \\ \zeta_\tau \times \mathbf{n} \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{d}^1 \zeta = \begin{pmatrix} -c_{\lambda,\mu}^0 \operatorname{div} \zeta_\tau \\ 0 \\ 0 \\ 0 \end{pmatrix}, \\
\mathbf{d}^2 \zeta &= \begin{pmatrix} \cdot \\ \cdot \\ \zeta_s \\ \cdot \end{pmatrix}, \quad \mathbf{d}^3 \zeta = \begin{pmatrix} \cdot \\ \cdot \\ 0 \\ -c_{\lambda,\mu}^0 \Delta_\tau \zeta_s \end{pmatrix}
\end{aligned} \quad (24)$$

where $c_{\lambda,\mu}^0$ is the constant in (18), $c_{\lambda,\mu}^m$ is another positive constant and \cdot indicates the presence of higher order operators on ζ_τ .

Note that the first three equations in (21): $\mathbf{L}^0 \zeta^{-2} = 0$, $\mathbf{L}^0 \zeta^{-1} = 0$, $\mathbf{L}^0 \zeta^0 + \mathbf{L}^2 \zeta^{-2} = \mathbf{R}^0 \mathbf{f}^0$ on S and $\mathbf{d}^0 \zeta^{-2} = 0$, $\mathbf{d}^0 \zeta^{-1} + \mathbf{d}^1 \zeta^{-2} = 0$, $\mathbf{d}^0 \zeta^0 + \mathbf{d}^1 \zeta^{-1} + \mathbf{d}^2 \zeta^{-2} = 0$ on ∂S , give back (15), (18), and (19) together with the fact that $\zeta_\tau^{-2} = \zeta_\tau^{-1} = 0$.

2.2.5 Boundary layer terms

The terms \mathbf{w}^k have a quite different structure. Their natural variables are (R, s, X_3) , see Section 2.1 and Fig. 3.



Figure 3. Boundary layer coordinates in $\delta S \times \Sigma_+$.

and they are easier to describe in boundary fitted components (w_r, w_s, w_3) corresponding to the local coordinates (r, s, x_3) . The first boundary layer term, \mathbf{w}^0 is a bending displacement in the sense of (5) and has a tensor product form: In boundary fitted components it reads

$$\begin{aligned}
w_s^0 &= 0 \quad \text{and} \quad (w_r^0, w_3^0)(R, s, X_3) = \varphi(s) \bar{w}_r^0(R, X_3) \\
\text{with} \quad \varphi &= \Delta_\tau \zeta_3^{-1}|_{\partial S}
\end{aligned}$$

and \bar{w}_r^0 is a two component exponentially decreasing profile on the semi-strip $\Sigma_+ := \{(R, X_3), R > 0, |X_3| < 1\}$: There exists $\eta > 0$ such that

$$|e^{\eta R} \bar{w}_r^0(R, X_3)| \text{ is bounded as } R \rightarrow \infty$$

The least upper bound of such η is the smallest exponent η_0 arising from the Papkovitch-Fadle eigenfunctions; see Gregory and Wan (1984). Both components of \bar{w}_r^0 are nonzero.

The next boundary layer terms \mathbf{w}^k are combinations of products of (smooth) traces on δS by profiles $\bar{w}^{k,\ell}$ in (R, X_3) . These profiles have singularities at the corners $(0, \pm 1)$ of Σ_+ , according to the general theory of Kondrat'ev (1967). Thus, in contrast with the 'regular' terms \mathbf{v}^k , which are smooth up to the boundary of Ω , the terms \mathbf{w}^k do have singular parts along the edges $\delta S \times \{\pm 1\}$ of the plate. Finally, the edge singularities of the solution \mathbf{u}^ε of problem (3) are related with the boundary layer terms only; see Dauge and Gruais (1998a) for further details.

2.3 Properties of the displacement expansion outside the boundary layer

Let S' be a subset of S such that the distance between $\delta S'$ and δS is positive. As a consequence of expansion (6) there holds

$$\begin{aligned}
\mathbf{u}^\varepsilon(\mathbf{x}) &= \sum_{k=-2}^K \varepsilon^k \mathbf{v}^k(\mathbf{x}_\tau, X_3) + O(\varepsilon^{K+1}) \\
&\quad \text{uniformly for } \mathbf{x} \in \bar{S}' \times (-\varepsilon, \varepsilon)
\end{aligned}$$

Coming back to physical variables (\mathbf{x}_τ, X_3) , the expansion terms \mathbf{v}^k being polynomials of degree $k+2$ in X_3 (Lemma 1), we find that

$$\begin{aligned}
\mathbf{u}^\varepsilon(\mathbf{x}) &= \sum_{k=-2}^K \varepsilon^k \bar{\mathbf{v}}^{K,k}(\mathbf{x}_\tau, X_3) + O(\varepsilon^{K+1}) \\
&\quad \text{uniformly for } \mathbf{x} \in \bar{S}' \times (-\varepsilon, \varepsilon)
\end{aligned}$$

with fields $\tilde{v}^{K,k}$ being polynomials in x_3 of degree $K - k$. This means that the expansion (6) can also be seen as a *Taylor expansion* at the midsurface, provided we are at a fixed positive distance from the lateral boundary.

Let us write the first terms in the expansions of the bending and membrane parts u_b^e and u_m^e of u^e :

$$u_b^e = \varepsilon^{-2} \left(-x_3 \nabla_\tau \zeta_3^{-2} + \zeta_3^{-2} + \frac{\lambda x_3^2}{2\lambda + 4\mu} \Delta_\tau \zeta_3^{-2} \right) - \left(0, \frac{\lambda}{6\lambda + 12\mu} \Delta_\tau \zeta_3^{-2} \right) + \varepsilon^{-1} \left(-x_3 \nabla_\tau \zeta_3^{-1}, \zeta_3^{-1} + \frac{\lambda x_3^2}{2\lambda + 4\mu} \Delta_\tau \zeta_3^{-1} \right) + \dots \quad (25)$$

From this formula, we can deduce the following asymptotics for the strain and stress components

$$\begin{aligned} e_{33}(u_b^e) &= -\varepsilon^{-2} x_3 \partial_{33}(\zeta_3^{-2} + \varepsilon \zeta_3^{-1}) + O(\varepsilon) \\ e_{33}(u_m^e) &= \varepsilon^{-2} \frac{\lambda x_3}{\lambda + 2\mu} \Delta_\tau (\zeta_3^{-2} + \varepsilon \zeta_3^{-1}) + O(\varepsilon) \quad (26) \\ \sigma^{33}(u_b^e) &= O(\varepsilon) \end{aligned}$$

Since $\varepsilon^{-2} x_3 = O(\varepsilon^{-1})$, we see that $e_{33} = O(\varepsilon^{-1})$. Thus, σ^{33} is two orders of magnitude less than e_{33} , which means a *plane stress limit*. To compute the shear strain (or stress), we use one further term in the asymptotics of u_b^e and obtain that it is one order of magnitude less than e_{33} :

$$e_{33}(u_m^e) = \frac{2\lambda + 2\mu}{\lambda + 2\mu} (\varepsilon^{-2} x_3^2 - 1) \partial_a \Delta_\tau \zeta_3^{-2} + O(\varepsilon) \quad (27)$$

Computations for the membrane part u_m^e are simpler and yield similar results

$$\begin{aligned} u_m^e &= \left(\xi_\tau^0, -\frac{\lambda x_3}{\lambda + 2\mu} \operatorname{div} \xi_\tau^0 \right) + \varepsilon \left(\xi_\tau^1, -\frac{\lambda x_3}{\lambda + 2\mu} \operatorname{div} \xi_\tau^1 \right) + \dots \\ e_{33}(u_m^e) &= \frac{1}{2} (\partial_a \xi_\tau^0 + \partial_b \xi_\tau^0) + \frac{\varepsilon}{2} (\partial_a \xi_\tau^1 + \partial_b \xi_\tau^1) + O(\varepsilon^2) \\ e_{33}(u_m^e) &= -\frac{\lambda}{\lambda + 2\mu} \operatorname{div} (\xi_\tau^0 + \varepsilon \xi_\tau^1) + O(\varepsilon^2) \quad (28) \end{aligned}$$

and $\sigma^{33}(u_m^e) = O(\varepsilon^2)$, $e_{33}(u_m^e) = O(\varepsilon)$.

In (26)–(28) the $O(\varepsilon)$ and $O(\varepsilon^2)$ are uniform on any region $\tilde{S} \times (-\varepsilon, \varepsilon)$ where the boundary layer terms have no influence. We postpone global energy estimates to the next section.

2.4 Eigen-mode problem

For each $\varepsilon > 0$, the spectrum of problem (4) is discrete and positive. Let $\Lambda_{b,j}^e$, $j = 1, 2, \dots$ be the increasing sequence of eigenvalues. In Ciarlet and Kesavan (1981) it is proved that $\varepsilon^{-2} \Lambda_{b,j}^e$ converges to the j th eigenvalue $\Lambda_{b,j}^K$ of the Dirichlet problem for the Kirchhoff operator L_b , cf. (16). In Nazarov and Zorin (1989) and Nazarov (1991c), a two-term asymptotics is constructed for the $\varepsilon^{-2} \Lambda_{b,j}^e$. Nazarov (2000b) proves that $|\varepsilon^{-2} \Lambda_{b,j}^e - \Lambda_{b,j}^K|$ is bounded by an $O(\sqrt{\varepsilon})$ for a much more general material matrix A .

In Dauge *et al.* (1999), full asymptotic expansions for eigenvalues and eigenvectors are proved: For each j there exist

- bending generators $\zeta_3^{-2}, \zeta_3^{-1}, \dots$ where ζ_3^{-2} is an eigen-vector of L_b associated with $\Lambda_{b,j}^K$
- real numbers $\Lambda_{b,j}^1, \Lambda_{b,j}^2, \dots$
- eigenvectors $u_{b,j}^e$ associated with $\Lambda_{b,j}^e$ for any $\varepsilon \in (0, \varepsilon_0)$

so that for any $K \geq 0$

$$\begin{aligned} \Lambda_{b,j}^e &= \varepsilon^2 \Lambda_{b,j}^K + \varepsilon^3 \Lambda_{b,j}^1 + \dots + \varepsilon^{K+2} \Lambda_{b,j}^K + O(\varepsilon^{K+3}) \\ u_{b,j}^e &= \varepsilon^{-2} (-x_3 \nabla_\tau \zeta_3^{-2}, \zeta_3^{-2}) + \varepsilon^{-1} (-x_3 \nabla_\tau \zeta_3^{-1}, \zeta_3^{-1}) \\ &\quad + \dots + \varepsilon^K (v^K + \chi w^K) + O(\varepsilon^{K+1}) \quad (29) \end{aligned}$$

where the terms v^K and w^K are generated by the ζ_3^k , $k \geq 0$ in a similar way as in Section 2.2, and $O(\varepsilon^{K+1})$ is uniform over Ω^e .

The bending and membrane displacements are the eigenvectors of the symmetry operator \mathcal{S} ; see (5). Since \mathcal{S} commutes with the elasticity operator, both have a joint spectrum, which means that there exists a basis of common eigenvectors. In other words, each elasticity eigenvalue can be identified as a bending or a membrane eigenvalue. The expansion (29) is the expansion of *bending* eigen-pairs.

The expansion of membrane eigen-pairs can be done in a similar way. Let us denote by $\Lambda_{m,j}^K$ the j th membrane eigenvalue on Ω^e and by $\Lambda_{m,j}^K$ the j th eigenvalue of the plane stress operator L_m , cf. (20) with Dirichlet boundary conditions. Then we have a similar statement as above, with the distinctive feature that the membrane eigenvalues tend to those of the plane stress model:

$$\Lambda_{m,j}^e = \Lambda_{m,j}^K + \varepsilon^3 \Lambda_{m,j}^1 + \dots + \varepsilon^K \Lambda_{m,j}^K + O(\varepsilon^{K+1}) \quad (30)$$

This fact, compared with (29), explains why the smallest eigenvalues are bending. Note that the eigenvalue formal series $\Lambda[s]$ satisfy reduced equations $L[s]\xi[s] = \Lambda[s]\xi[s]$ like (21) with the same L^0 , $L^1 = 0$ and L^2 as in (22). In

particular, equations

$$\begin{pmatrix} L_m & 0 \\ 0 & \varepsilon^2 L_b \end{pmatrix} \begin{pmatrix} \xi_\tau \\ \zeta_3 \end{pmatrix} = \Lambda \begin{pmatrix} \xi_\tau \\ \zeta_3 \end{pmatrix} \quad (31)$$

give back the ‘limiting’ eigenvalues $\Lambda_{m,j}^K$ and $\varepsilon^2 \Lambda_{b,j}^K$. Our last remark is that the second terms $\Lambda_{b,j}^1$ and $\Lambda_{m,j}^1$ are positive; see Dauge and Yosibash (2002) for a discussion of that fact.

2.5 Extensions

2.5.1 Traction on the free parts of the boundary

Instead of a volume load, or in addition to it, tractions g^\pm can be imposed on the faces $S \times \{\pm\varepsilon\}$ of the plate. Let us assume that g^\pm is independent of ε . Then the displacement u^e has a similar expansion as in (6), with the following modifications:

- If the bending part of g^\pm is nonzero, then the regular part starts with $\varepsilon^{-3} v^{-3}$ and the boundary layer part with $\varepsilon^{-1} \chi w^{-1}$;
- If the membrane part of g^\pm is nonzero, the membrane regular part starts with $\varepsilon^{-1} v^{-1}$.

2.5.2 Lateral boundary conditions

A similar analysis holds for each of the seven remaining types of ‘canonical’ boundary conditions: soft clamping, hard simple support, soft simple support, two types of friction, sliding edge, and free boundary. See Dauge, Gruais and Rösle (1999/00) for details. It would also be possible to extend such an analysis to more intimately mixed boundary conditions where only moments through the thickness along the lateral boundary are imposed for displacement or traction components; see Schwab (1996).

If, instead of volume load f or tractions g^\pm , we set $f = 0$, $g^\pm = 0$, and impose nonzero lateral boundary conditions, u^e will have a similar expansion as in (6) with the remarkable feature that the degree of the regular part in the thickness variable is ≤ 3 ; see Dauge and Schwab (2002), Rem. 5.4. Moreover, in the clamped situation, the expansion starts with $O(1)$.

2.5.3 Laminated composites

If the material of the plate is homogeneous, but not isotropic, u^e will still have a similar expansion; see Dauge and Gruais (1996) and Dauge and Yosibash (2002) for orthotropic plates. If the plate is laminated, that is, formed by the union of several plies made of different homogeneous materials, then u^e still expands in regular parts v^k and

boundary layer parts w^k , but the v^k are no more polynomials in the thickness variable, only *piecewise polynomial* in each ply, and continuous; see Actis, Szabo and Schwab (1999). Nazarov (2000a, 2000b) addresses more general material laws where the matrix A depends on the variables x_τ and $x_3 = x_3/\varepsilon$.

3 HIERARCHICAL MODELS FOR PLATES

3.1 The concepts of hierarchical models

The idea of hierarchical models is a natural and efficient extension to that of limiting models and dimension reduction. In the finite element framework, it has been firstly formulated in Szabó and Sahrman (1988) for isotropic domains, mathematically investigated in Babuška and Li (1991, 1992a, 1992b), and generalized to laminated composites in Babuška, Szabó and Actis (1992) and Actis, Szabo and Schwab (1999). A hierarchy of models consists of

- a sequence of subspaces $V^q(\Omega^e)$ of $V(\Omega^e)$ with the orders $q = (q_1, q_2, q_3)$ forming a sequence of integer triples, satisfying

$$V^q(\Omega^e) \subset V^{q'}(\Omega^e) \quad \text{if} \quad q \leq q' \quad (32)$$

- a sequence of related Hooke laws $\sigma = A_q \varepsilon$, corresponding to a sequence of elastic bilinear forms $a^{k,q}(u, u') = \int_{\Omega^e} A_q \varepsilon(u) : \varepsilon(u')$.

Let $u^{k,q}$ be the solution of the problem

$$\text{Find } u^{k,q} \in V^q(\Omega^e) \text{ such that} \\ a^{k,q}(u^{k,q}, u') = \int_{\Omega^e} f \cdot u' \, dx, \quad \forall u' \in V^q(\Omega^e) \quad (33)$$

Note that problem (33) is a Galerkin projection of problem (3) if $a^{k,q} = a^k$.

Any model that belongs to the hierarchical family has to satisfy three requirements; see Szabó and Babuška (1991), Chap. 14.5:

- (a) *Approximability*. At any fixed thickness $\varepsilon > 0$:

$$\lim_{q \rightarrow \infty} \|u^k - u^{k,q}\|_{E(\Omega^e)} = 0 \quad (34)$$

- (b) *Asymptotic consistency*. For any fixed degree q :

$$\lim_{\varepsilon \rightarrow 0} \frac{\|u^k - u^{k,q}\|_{E(\Omega^e)}}{\|u^k\|_{E(\Omega^e)}} = 0 \quad (35)$$

(c) *Optimality of the convergence rate.* There exists a sequence of positive exponents $\gamma(q)$ with the growth property $\gamma(q) < \gamma(q')$ if $q < q'$, such that 'in the absence of boundary layers and edge singularities':

$$\|u^\varepsilon - u^{\varepsilon,0}\|_{E(\Omega^*)} \leq C\varepsilon^{\gamma(q)} \|u^\varepsilon\|_{E(\Omega^*)} \quad (36)$$

The substantiation of hierarchical models for plates, in general, requires the choice of three sequences of finite dimensional nested director spaces $\Psi_0^j \subset \dots \subset \Psi_N^j \subset \dots \subset H^1(-1, 1)$ for $j = 1, 2, 3$ and the definition of the space $V^q(\Omega^*)$ for $q = (q_1, q_2, q_3)$ as

$$V^q(\Omega^*) = \left\{ u \in V(\Omega^*), (x_\tau, x_3) \mapsto u_j(x_\tau, \varepsilon x_3) \in H_0^1(S) \otimes \Psi_j^q, \quad j = 1, 2, 3 \right\} \quad (37)$$

We can reformulate (37) with the help of director functions: With $d_j(N)$ being the dimension of Ψ_j^N , let $\Phi_j^N = \Phi_j^N(x_3)$, $0 \leq n \leq d_j(N)$, be hierarchic bases (the director functions) of Ψ_j^N . There holds

$$V^q(\Omega^*) = \left\{ u \in V(\Omega^*), \exists \sigma_j^n \in H_0^1(S), 0 \leq n \leq d_j(q_j), u_j(x_\tau, x_3) = \sum_{n=0}^{d_j(q_j)} \sigma_j^n(x_\tau) \Phi_j^n\left(\frac{x_3}{\varepsilon}\right) \right\} \quad (38)$$

The choice of the best director functions is addressed in Vogelius and Babuška (1981c) in the case of second-order scalar problems with general coefficients (including possible stratifications). For smooth coefficients, the space Ψ_j^N coincides with the space \mathcal{P}_N of polynomial with degree $\leq N$. The director functions can be chosen as the Legendre polynomials $L_n(x_3)$ or, simply, the monomials X_3^n (and then X_3^n can be used equivalently instead of $(x_3/\varepsilon)^n$ in (38)).

We describe in the sequel in more detail, the convenient hierarchies for plates and discuss the three qualities (34)–(36); see Babuška and Li (1991, 1992a) and Babuška, Szabó and Actis (1992) for early references.

3.2 The limit model (Kirchhoff–Love)

In view of expansion (6), we observe that if the transverse component f_3 of the load is nonzero on the midsurface, u^ε is unbounded as $\varepsilon \rightarrow 0$. If we multiply by ε^2 , we have a convergence to $(0, \zeta_\tau^{-2})$, which is not kinematically relevant. At that level, a correct notion of limit uses scalings of coordinates: If we define the scaled displacement \tilde{u}^ε by its components on the stretched plate $\Omega = S \times (-1, 1)$ by

$$\tilde{u}_\tau := \varepsilon u_\tau^\varepsilon \quad \text{and} \quad \tilde{u}_3 := \varepsilon^2 u_3^\varepsilon \quad (39)$$

then \tilde{u}^ε converges to $(-X_3 \nabla_\tau \zeta_3^{-2}, \zeta_3^{-2})$ in $H^1(\Omega)^3$ as $\varepsilon \rightarrow 0$. This result, together with the mathematical derivation of the resultant equation (15), is due to Ciarlet and Destuynder (1979a).

The corresponding subspace of $V(\Omega^*)$ is that of bending Kirchhoff displacements or, more generally, of Kirchhoff displacements:

$$V^{KL}(\Omega^*) = \{ u \in V(\Omega^*), \exists \zeta \in H_0^1(S) \times H_0^1(S), u = (\zeta_\tau - X_3 \nabla_\tau \zeta_3, \zeta_3) \} \quad (40)$$

It follows from (40) that $e_{13} = e_{23} = 0$ for which the physical interpretation is that 'normals to S prior to deformation remain straight lines and normals after deformation'. Hooke's law has to be modified with the help of what we call 'the plane stress trick'. It is based on the assumption that the component σ^{33} of the stress is negligible (note that the asymptotics (6) of the three-dimensional solution yields that $\sigma^{33} = O(\varepsilon)$, whereas $e_{33} = O(\varepsilon^{-1})$ outside the boundary layer, cf. (26), which justifies the plane stress assumption). From standard Hooke's law (1), we extract the relation $\sigma^{33} = \lambda(e_{11} + e_{22}) + (\lambda + 2\mu)e_{33}$, then set σ^{33} to zero, which yields

$$e_{33} = -\frac{\lambda}{\lambda + 2\mu} (e_{11} + e_{22}) \quad (41)$$

Then, we modify Hooke's law (1) by substituting e_{33} by its expression (41) in σ^{11} and σ^{22} , to obtain

$$\sigma^{ii} = \frac{2\lambda\mu}{\lambda + 2\mu} (e_{11} + e_{22}) + 2\mu e_{ii}, \quad i = 1, 2 \quad (42)$$

$$\sigma^{ij} = 2\mu e_{ij} \quad \text{for } i \neq j$$

Thus, $\tilde{\sigma}^{ii} = \tilde{\lambda}(e_{11} + e_{22}) + 2\mu e_{ii}$, with $\tilde{\lambda}$ given by (17). Taking into account that $e_{33} = 0$ for the elements of $V^{KL}(\Omega^*)$, we obtain a new Hooke's law given by the same formulae as (1) when replacing the Lamé coefficient λ by $\tilde{\lambda}$. This corresponds to a modified material matrix \tilde{A}^{ijkl}

$$\tilde{A}^{ijkl} = \tilde{\lambda} \delta^{ij} \delta^{kl} + \mu (\delta^{ik} \delta^{jl} + \delta^{il} \delta^{jk}) \quad (43)$$

and a reduced elastic energy $\tilde{a}(u, u) = \int_{\Omega^*} \sigma^{ij}(u) e_{ij}(u)$. Note that for $u = (\zeta_\tau - X_3 \nabla_\tau \zeta_3, \zeta_3)$

$$\tilde{a}(u, u) = 2\varepsilon \int_S \tilde{A}^{\alpha\beta\alpha\beta} e_{\alpha\beta}(\zeta_\tau) e_{\alpha\beta}(\zeta_\tau) dx_\tau + \frac{2\varepsilon^3}{3} \int_S \tilde{A}^{\alpha\beta\alpha\beta} e_{\alpha\beta}(\zeta_3) e_{\alpha\beta}(\zeta_3) dx_\tau \quad (44)$$

exhibiting a *membrane part* in $O(\varepsilon)$ and a *bending part* in $O(\varepsilon^3)$. There hold as a consequence of Theorem 1

Theorem 2. Let $u^{\varepsilon, KL}$ be the solution of problem (33) with $V^q = V^{KL}$ and $a^q = \tilde{a}$. Then

- In general $u^{\varepsilon, KL} = \varepsilon^{-2} (-X_3 \nabla_\tau \zeta_3^{-2}, \zeta_3^{-2}) + O(1)$ with ζ_3^{-2} the solution of (15);
- If f is membrane, $u^{\varepsilon, KL} = (\zeta_\tau^0, 0) + O(\varepsilon^2)$ with ζ_τ^0 the solution of (19).

Can we deduce the asymptotic consistency for that model? No! Computing the lower-order terms in the expression (35), we find with the help of (25) that, if $f_3 \neq 0$

$$\|u^\varepsilon\|_{E(\Omega^*)} \simeq O(\varepsilon^{-1/2})$$

and

$$\|u^\varepsilon - u^{\varepsilon, KL}\|_{E(\Omega^*)} \geq \|e_{33}(u^\varepsilon)\|_{L^2(\Omega^*)} \simeq O(\varepsilon^{-1/2})$$

Another source of difficulty is that, eventually, relation (41) is not satisfied by $u^{\varepsilon, KL}$. If $f_3^0 = 0$ and $f_3^0 \neq 0$, we have exactly the same difficulties with the membrane part.

A way to overcome these difficulties is to consider a complementing operator C defined on the elements of V^{KL} by

$$Cu = u + \left(0, -\frac{\lambda}{\lambda + 2\mu} \int_0^2 \operatorname{div} u_\tau(\cdot, y) dy \right) \quad (45)$$

Then (41) is now satisfied by Cu for any $u \in V^{KL}$. Moreover (still assuming $f_3 \neq 0$), one can show

$$\|u^\varepsilon - Cu^{\varepsilon, KL}\|_{E(\Omega^*)} \leq C\sqrt{\varepsilon} \|u^\varepsilon\|_{E(\Omega^*)} \quad (46)$$

The error factor $\sqrt{\varepsilon}$ is due to the first boundary layer term w^0 . The presence of w^0 is a direct consequence of the fact that $Cu^{\varepsilon, KL}$ does not satisfy the lateral boundary conditions.

Although the Kirchhoff–Love model is not a member of the hierarchic family, it is the limit of all models for $\varepsilon \rightarrow 0$.

3.3 The Reissner–Mindlin model

This model is obtained by enriching the space of kinematically admissible displacements, allowing normals to S to rotate after deformation. Instead of (40), we set

$$V^{RM}(\Omega^*) = \{ u \in V(\Omega^*), \exists \zeta \in H_0^1(S)^3, \exists \theta_\tau \in H_0^1(S)^2, u = (z_\tau - X_3 \theta_\tau, z_3) \}$$

With the elasticity tensor A corresponding to 3-D elasticity, the displacements and strain–energy limit of the RM model as $d \rightarrow 0$ would not coincide with the 3-D limit (or the Kirchhoff–Love limit).

We have again to use instead the reduced elastic bilinear form \tilde{a} to restore the convergence to the correct limit, by virtue of the same plane stress trick. The corresponding elasticity tensor is \tilde{A} (43). A further correction can be introduced in the shear components of \tilde{A} to better represent the fully 3-D shear stresses σ^{13} and σ^{23} (and also the strain energy) for small yet nonzero thickness ε . The material matrix entries A^{1313} , A^{2323} are changed by introducing the so-called shear correction factor κ :

$$\tilde{A}^{1313} = \kappa A^{1313} \quad \tilde{A}^{2323} = \kappa A^{2323}$$

By properly chosen κ , either the energy of the RM solution, or the deflection u_3 can be optimized with respect to the fully 3-D plate. The smaller the ε , the smaller the influence of κ on the results. For the isotropic case, two possible κ 's are (see details in Babuška, d'Harcourt and Schwab (1991a)):

$$\kappa_{\text{Bergy}} = \frac{5}{6(1-\nu)} \quad \text{or} \quad \kappa_{\text{Deflection}} = \frac{20}{3(8-3\nu)},$$

with $\nu = \frac{\lambda}{2(\lambda + \mu)}$ (Poisson ratio)

A value of $\kappa = 5/6$ is frequently used in engineering practice, but for modal analysis, no optimal value of κ is available.

Note that, by integrating equations of (33) through the thickness, we find that problem (33) is equivalent to a variational problem for z and θ only. For the elastic energy, we have

$$\begin{aligned} \tilde{a}(u, u) &= 2\varepsilon \int_S \tilde{A}^{\alpha\beta\alpha\beta} e_{\alpha\beta}(z_\tau) e_{\alpha\beta}(z_\tau) dx_\tau \\ &\quad + \varepsilon \int_S \kappa \mu (\partial_\alpha z_3 - \theta_\alpha) (\partial_\alpha z_3 - \theta_\alpha) dx_\tau \\ &\quad + \frac{2\varepsilon^3}{3} \int_S \tilde{A}^{\alpha\beta\alpha\beta} e_{\alpha\beta}(\theta_\tau) e_{\alpha\beta}(\theta_\tau) dx_\tau \end{aligned} \quad (47)$$

(membrane energy)
(shear energy)
(bending energy)

Let $u^{\varepsilon, RM}$ be the solution of problem (33) with $V^q = V^{RM}$ and $a^q = \tilde{a}$. The singular perturbation character appears clearly. In contrast with the Kirchhoff–Love model, the solution admits a boundary layer part. Arnold and Falk (1990b, 1996) have described the two-scale asymptotics of $u^{\varepsilon, RM}$. Despite the presence of boundary layer terms, the question of knowing if $u^{\varepsilon, RM}$ is closer to u^ε than $u^{\varepsilon, KL}$ has no clear answer to our knowledge. A careful investigation of the first eigenvalues Λ_1^ε , $\Lambda_1^{\varepsilon, KL}$, and $\Lambda_1^{\varepsilon, RM}$ of these three models in the case of lateral Dirichlet conditions shows the following behavior for ε small enough (Dauge and Yosibash, 2002):

$$\Lambda_1^{\varepsilon, \text{RM}} < \Lambda_1^{\varepsilon, \text{KL}} < \Lambda_1^{\varepsilon}$$

which tends to prove that RM model is not generically better than KL for (very) thin plates. Nevertheless, an estimate by the same asymptotic bound as in (46) is valid for $u^\varepsilon - Cu^\varepsilon, \text{RM}$.

3.4 Higher order models

The RM model is a (1, 1, 0) model with reduced elastic energy. For any $\mathbf{q} = (q_1, q_2, q_3)$ we define the space $V^{\mathbf{q}}$ by (compare with (38) for monomial director functions)

$$V^{\mathbf{q}}(\Omega^\varepsilon) = \left\{ u \in V(\Omega^\varepsilon), \quad \exists \varphi_n^\varepsilon \in H_0^1(S)^2, \quad 0 \leq n \leq q_1, \right. \\ \left. \exists \varphi_n^\varepsilon \in H_0^1(S), \quad 0 \leq n \leq q_2, \right. \\ \left. u_\tau = \sum_{n=0}^{q_1} x_3^n \varphi_n^\varepsilon(x_\tau) \quad \text{and} \quad u_3 = \sum_{n=0}^{q_2} x_3^n \varphi_n^\varepsilon(x_\tau) \right\} \quad (48)$$

The subspaces $V_0^{\mathbf{q}}$ and $V_n^{\mathbf{q}}$ of bending and membrane displacements in $V^{\mathbf{q}}$ can also be used, according to the nature of the data. The standard 3-D elastic energy (2) is used with $V^{\mathbf{q}}$ and $V_0^{\mathbf{q}}$ for any $\mathbf{q} \geq (1, 1, 2)$ and with $V_n^{\mathbf{q}}$ for any $\mathbf{q} \geq (0, 0, 1)$.

Theorem 3.

- (i) If \mathbf{f} satisfies $\mathbf{f}_3|_S \neq 0$, for any $\mathbf{q} \geq (1, 1, 2)$ there exists $C_q = C_q(\mathbf{f}) > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$

$$\|u^\varepsilon - u^{\varepsilon, \mathbf{q}}\|_{E(\Omega^\varepsilon)} \leq C_q \sqrt{\varepsilon} \|u^\varepsilon\|_{E(\Omega^\varepsilon)} \quad (49)$$

- (ii) If \mathbf{f} is membrane and $\mathbf{f}_\tau|_S \neq 0$, for any $\mathbf{q} \geq (0, 0, 1)$ there exists $C_q = C_q(\mathbf{f}) > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$ (49) holds.

Proof. Since the energy is not altered by the model, $u^{\varepsilon, \mathbf{q}}$ is a Galerkin projection of u^ε on $V^{\mathbf{q}}(\Omega^\varepsilon)$. Since the strain energy is uniformly equivalent to the elastic energy on any Ω^ε , we have by Céa's lemma that there exists $C > 0$

$$\|u^\varepsilon - u^{\varepsilon, \mathbf{q}}\|_{E(\Omega^\varepsilon)} \leq C \|u^\varepsilon - v^{\mathbf{q}}\|_{E(\Omega^\varepsilon)} \quad \forall v^{\mathbf{q}} \in V^{\mathbf{q}}(\Omega^\varepsilon)$$

- (i) We choose, compare with (25),

$$v^{\mathbf{q}} = \varepsilon^{-2} \left(-x_3 \nabla_\tau \zeta_3^{-2}, \quad \zeta_3^{-2} + \frac{\lambda x_3^2}{2\lambda + 4\mu} \Delta_\tau \zeta_3^{-2} \right) \\ - \frac{\varepsilon^{-2} \lambda x_3^2}{2\lambda + 4\mu} \varphi(s) (\theta, \xi(R))$$

with $\varphi = \Delta_\tau \zeta_3^{-2}|_{\partial S}$ and ξ a smooth cut-off function equal to 1 in a neighborhood of $R = 0$ and 0 for $R \geq 1$. Then $v^{\mathbf{q}}$ satisfies the lateral boundary conditions and we can check (49) by combining Theorem 1 with the use of Céa's lemma.

- (ii) We choose, instead

$$v^{\mathbf{q}} = \left(\zeta_\tau^0, -\frac{\lambda x_3}{\lambda + 2\mu} \operatorname{div} \zeta_\tau^0 \right) + \frac{\lambda x_3}{\lambda + 2\mu} \varphi(s) (\theta, \xi(R)) \\ \text{with } \varphi = \operatorname{div} \zeta_\tau^0|_{\partial S} \quad \square$$

It is worthwhile to mention that for the (1, 1, 2) model the shear correction factor (when $\nu \rightarrow 0$, $\kappa_{(1,1,2)}$ tends to 5/6, just like for the two shear correction factors of the RM model)

$$\kappa_{(1,1,2)} = \frac{12 - 2\nu}{\nu^2} \left(-1 + \sqrt{1 + \frac{20\nu^2}{(12 - 2\nu)^2}} \right)$$

can be used for optimal results in respect with the error in energy norm and deflection for finite thickness plates; see Babuška, d'Harcourt and Schwab (1991a). For higher plate models, no shear correction factor is furthermore needed.

The result in Schwab and Wright (1995) regarding the approximability of the boundary layers by elements of $V^{\mathbf{q}}$, yields that the constant C_q in (49) should rapidly decrease when \mathbf{q} increases. Nevertheless the factor $\sqrt{\varepsilon}$ is still present, for any \mathbf{q} , because of the presence of the boundary layer terms. The numerical experiments in Dauge and Yosibash (2000) demonstrate that the higher the degree of the hierarchical model, the better the boundary layer terms are approximated.

If one wants to have an approximation at a higher order in ε one should

- either consider a problem without boundary layer, as mentioned in requirement (c) (36), that is, a rectangular plate with symmetry boundary conditions: In this case, the convergence rate $\gamma(\mathbf{q})$ in ε is at least $\min_j q_j - 1$,
- or combine a hierarchy of models with a three-dimensional discretization of the boundary layer; see Stein and Ohnibus (1969) and Dauge and Schwab (2002).

The (1, 1, 2) is the lowest order model which is asymptotically consistent for bending. See Paumier and Raoult (1997) and Rösle *et al.* (1999). It is the first model in the bending model hierarchy

$$(1, 1, 2), \quad (3, 3, 2), \quad (3, 3, 4), \dots \\ (2n - 1, 2n - 1, 2n), \quad (2n + 1, 2n + 1, 2n), \dots$$

The exponent $\gamma(\mathbf{q})$ in (36) can be proved to be $2n - 1$ if $\mathbf{q} = (2n - 1, 2n - 1, 2n)$ and $2n$ if $\mathbf{q} = (2n + 1, 2n + 1, 2n)$, thanks to the structure of the operator series $V[\varepsilon]$ and $Q[\varepsilon]$ in (11). If the load \mathbf{f} is constant over the whole plate, then the model of degree (3, 3, 4) captures the whole regular part of u^ε , (Dauge and Schwab (2002), Rem. 8.3) and if, moreover, $\mathbf{f} \equiv 0$ (in this case, only a lateral boundary condition is imposed), the degree (3, 3, 2) is sufficient.

3.5 Laminated plates

If the plate is laminated, the material matrix $A = A^\varepsilon$ has a sandwich structure, depending on the thickness variable x_3 : We assume that $A^\varepsilon(x_3) = A(x_{-3})$, where the coefficients of A are piecewise constant. In Nazarov (2000a) the asymptotic analysis is started, including such a situation. We may presume that a full asymptotic expansion like (6) with a similar internal structure, is still valid.

In the homogeneous case, the director functions in (38) are simply the monomials of increasing degrees; see (48). In the laminated case, the first director functions are still 1 and x_3 :

$$\Phi_1^0 = \Phi_2^0 = \Phi_3^0 = 1; \quad \Phi_1^1 = \Phi_2^1 = x_3$$

In the homogeneous case, we have $\Phi_j^1 = x_j$ and $\Phi_j^2 = x_j^2$, $j = 1, 2, 3$. In Actis, Szabo and Schwab (1999) three more piecewise linear director functions and three piecewise quadratic director functions are exhibited for the laminated case.

How many independent director functions are necessary to increase the convergence rate $\gamma(\mathbf{q})$ (36)? In other words, what is the dimension of the spaces $\Psi_j^{\mathbf{q}}$ (cf. (37))? In our formalism, see (10)–(11), this question is equivalent to knowing the structure of the operators V^j . Comparing with Nazarov (2000a), we can expect that

$$V^1 \zeta = \left(-X_3 \nabla_\tau \zeta_3, \quad P_3^{1,1}(X_3) \partial_1 \zeta_1 + P_3^{1,2}(X_3) (\partial_1 \zeta_2 + \partial_2 \zeta_1) \right. \\ \left. + P_3^{1,3}(X_3) \partial_2 \zeta_2 \right) \\ V^2 \zeta = \left(\sum_{k=1}^3 P_j^{2,k,1}(X_3) \partial_k^2 \zeta_k + P_j^{2,k,2}(X_3) \partial_k^2 \zeta_k \right. \\ \left. + P_j^{2,k,3}(X_3) \partial_k^2 \zeta_k \right)_{j=1,2,3} \quad (50)$$

As soon as the above functions $P_j^{n,k}$ are independent, they should be present in the bases of the director space $\Psi_j^{\mathbf{q}}$. The dimensions of the spaces generated by the $P_j^{n,k}$ have upper bounds depending only on n . But their actual dimensions depend on the number of plies and their nature.

4 MULTISCALE EXPANSIONS AND LIMITING MODELS FOR SHELLS

Up to now, the only available results concerning multiscale expansions for 'true' shells concern the case of clamped elliptic shells investigated in Faou (2001a,b, 2003). For (physical) shallow shells, which are closer to plates than shells, multiscale expansions can also be proved; see Nazarov (2000a) and Andreou and Faou (2001).

In this section, we describe the results for clamped elliptic shells, then present the main features of the classification of shells as flexural and membrane. As a matter of fact, multiscale expansions are known for the most extreme representatives of the two types: (i) plates for flexural shells, (ii) clamped elliptic shells for membrane shells. Nevertheless, multiscale expansions in the general case seem out of reach (or, in certain cases, even irrelevant) (see Chapter 3, Volume 2).

4.1 Curvature of a midsurface and other important tensors

We introduce minimal geometric tools, namely, the metric and curvature tensors of the midsurface S , the change of metric tensor γ_{ab} , and the change of curvature tensor ρ_{ab} . We also address the essential notions of elliptic, hyperbolic, or parabolic point in a surface. We make these notions more explicit for axisymmetric surfaces. A general introduction to differential geometry on surfaces can be found in Stoker (1969).

Let us denote by $(X, Y)_{\mathbb{R}^3}$ the standard scalar product of two vectors X and Y in \mathbb{R}^3 . Using the fact that the midsurface S is embedded in \mathbb{R}^3 , we naturally define the metric tensor (a_{ab}) as the projection on S of the standard scalar product in \mathbb{R}^3 : Let \mathbf{p}_τ be a point of S and X, Y , two tangent vectors to S in \mathbf{p}_τ . In a coordinate system $\mathbf{x}_\tau = (x_a)$ on S , the components of X and Y are (X^τ) and (Y^τ) , respectively. Then the matrix $(a_{ab}(\mathbf{x}_\tau))$ is the only positive definite symmetric 2×2 matrix such that for all such vectors X and Y

$$(X, Y)_{\mathbb{R}^3} = a_{ab}(\mathbf{x}_\tau) X^\tau Y^\tau =: (X, Y)_S$$

The inverse of a_{ab} is written a^{ab} and thus satisfies $a^{ab} a_{bc} = \delta_c^a$, where δ_c^a is the Kronecker symbol and where we used the repeated indices convention for the contraction of tensors.

The covariant derivative D is associated with the metric a_{ab} as follows: It is the unique differential operator such that $D(X, Y)_S = (DX, Y)_S + (X, DY)_S$ for all vector fields X and Y . In a local coordinate system, we have

$$D_a = \partial_a + \text{terms of order } 0$$

where ∂_a is the derivative with respect to the coordinate x_a . The terms of order 0 do depend on the choice of the coordinate system and on the type of the tensor field on which D is applied. They involve the Christoffel symbols of S in the coordinate system (x_a) .

The principal curvatures at a given point $\mathbf{p}_T \in S$ can be seen as follows: We consider the family \mathcal{P} of planes P containing \mathbf{p}_T and orthogonal to the tangent plane to S at \mathbf{p}_T . For $P \in \mathcal{P}$, $P \cap S$ defines a curve in P and we denote by κ its signed curvature κ . The sign of κ is determined by the orientation of S . The principal curvatures κ_1 and κ_2 are the minimum and maximum of κ when $P \in \mathcal{P}$. The principal radii of curvature are $R_i := |\kappa_i|^{-1}$. The Gaussian curvature of S in \mathbf{p}_T is $K(\mathbf{p}_T) = \kappa_1 \kappa_2$.

A point \mathbf{p}_T is said to be elliptic if $K(\mathbf{p}_T) > 0$, hyperbolic if $K(\mathbf{p}_T) < 0$, parabolic if $K(\mathbf{p}_T) = 0$ but κ_1 or κ_2 is nonzero, and planar if $\kappa_1 = \kappa_2 = 0$. An elliptic shell is a shell whose midsurface is everywhere elliptic up to the boundary (similar definitions hold for hyperbolic and parabolic shells... and planar shells that are plates).

The curvature tensor is defined as follows: Let $\Psi: \mathbf{x}_T \mapsto \Psi(\mathbf{x}_T)$ be a parameterization of S in a neighborhood of a given point $\mathbf{p}_T \in S$ and $\mathbf{n}(\Psi(\mathbf{x}_T))$ be the normal to S in $\Psi(\mathbf{x}_T)$. The formula

$$b_{ab} := \left(\mathbf{n}(\Psi(\mathbf{x}_T)) \cdot \frac{\partial \Psi}{\partial x_a}(\mathbf{x}_T) \right)_{\mathbf{R}^3}$$

defines, in the coordinate system $\mathbf{x}_T = (x_a)$, the components of a covariant tensor field on S , which is called the curvature tensor.

The metric tensor yields diffeomorphisms between tensor spaces of different types (covariant and contravariant): We have, for example, $b_a^b = a^{ab} b_{ab}$. With these notations, we can show that in any coordinate system, the eigenvalues of b_a^b at a point \mathbf{p}_T are the principal curvatures at \mathbf{p}_T .

In the special case where S is an axisymmetric surface parametrized by

$$\Psi: (x_1, x_2) \mapsto (x_1 \cos x_2, x_1 \sin x_2, f(x_1)) \in \mathbb{R}^3 \quad (51)$$

where $x_1 \geq 0$ is the distance to the axis of symmetry, $x_2 \in [0, 2\pi[$ is the angle around the axis, and $f: \mathbb{R} \mapsto \mathbb{R}$ a smooth function, we compute directly that

$$(b_a^b) = \frac{1}{\sqrt{1+f'(x_1)^2}} \begin{pmatrix} f''(x_1) & 0 \\ 0 & f'(x_1) \\ 0 & x_1 \end{pmatrix}$$

whence

$$K = \frac{f'(x_1) f''(x_1)}{x_1 (1 + f'(x_1)^2)^{3/2}}$$

A deformation pattern is a three-component field $\xi = (\xi_a, \xi_3)$ where ξ_a is a surface displacement on S , and ξ_3 a function on S . The change of metric tensor $\gamma_{ab}(\xi)$ associated with the deformation pattern ξ has the following expression:

$$\gamma_{ab}(\xi) = \frac{1}{2} (D_a \xi_b + D_b \xi_a) - b_{ab} \xi_3 \quad (52)$$

The change of curvature tensor associated with ξ writes

$$\rho_{ab}(\xi) = D_a D_b \xi_3 - b_a^c D_b \xi_c + b_b^c D_a \xi_c + D_a b_b^c \xi_c \quad (53)$$

4.2 Clamped elliptic shells

The generic member Ω^ϵ of our family of shells is defined as

$$S \times (-\epsilon, \epsilon) \ni (\mathbf{p}_T, x_3) \longrightarrow \mathbf{p}_T + x_3 \mathbf{n}(\mathbf{p}_T) \in \Omega^\epsilon \subset \mathbb{R}^3 \quad (54)$$

where $\mathbf{n}(\mathbf{p}_T)$ is the normal to S at \mathbf{p}_T . Now three stretched variables are required (cf. Section 2.1 for plates):

$$X_3 = \frac{x_3}{\epsilon}, \quad R = \frac{r}{\epsilon} \quad \text{and} \quad T = \frac{t}{\sqrt{\epsilon}}$$

where (r, s) is a system of normal and tangential coordinates to ∂S in S .

4.2.1 Three-dimensional expansion

The solutions of the family of problems (3) have a three-scale asymptotic expansion in powers of $\epsilon^{1/2}$, with regular terms $\mathbf{v}^{k/2}$, boundary layer terms $\mathbf{w}^{k/2}$ of scale ϵ like for plates, and new boundary layer terms $\mathbf{q}^{k/2}$ of scale $\epsilon^{1/2}$.

Theorem 4. (Faou, 2003) For the solutions of problems (3), there exist regular terms $\mathbf{v}^{k/2}(\mathbf{x}_T, X_3)$, $k \geq 0$, boundary layer terms $\mathbf{q}^{k/2}(T, s, X_3)$, $k \geq 0$ and $\mathbf{w}^{k/2}(R, s, X_3)$, $k \geq 2$, such that

$$\mathbf{u}^\epsilon \simeq (\mathbf{v}^0 + \chi \mathbf{q}^0) + \epsilon^{1/2} (\mathbf{v}^{1/2} + \chi \mathbf{q}^{1/2}) + \epsilon (\mathbf{v}^1 + \chi \mathbf{q}^1 + \chi \mathbf{w}^1) + \dots \quad (55)$$

in the sense of asymptotic expansions: There holds the following estimates

$$\left\| \mathbf{u}^\epsilon - \sum_{k=0}^{2K} \epsilon^{k/2} (\mathbf{v}^{k/2} + \chi \mathbf{q}^{k/2} + \chi \mathbf{w}^{k/2}) \right\|_{B(\Omega^\epsilon)} \leq C_K(\mathbf{T}) \epsilon^{K+1/2}, \quad K = 0, 1, \dots$$

where we have set $\mathbf{w}^0 = \mathbf{w}^{1/2} = 0$ and the constant $C_K(\mathbf{T})$ is independent of $\epsilon \in (0, \epsilon_0]$.

Like for plates, the terms of the expansion are linked with each other, and are generated by a series of deformation patterns $\xi^{k/2} = \xi^{k/2}(\mathbf{x}_T)$ of the midsurface S . They solve a recursive system of equations, which can be written in a condensed form as an equality between formal series, like for plates. The distinction from plates is that, now, half-integer powers of ϵ are involved and we write, for example, $\xi[\epsilon^{1/2}]$ for the formal series $\sum_k \epsilon^{k/2} \xi^{k/2}$.

4.2.2 Regular terms

The regular terms series $\mathbf{v}[\epsilon^{1/2}] = \sum_k \epsilon^{k/2} \mathbf{v}^{k/2}$ is determined by an equation similar to (10):

$$\mathbf{v}[\epsilon^{1/2}] = \mathbf{V}[\epsilon^{1/2}] \xi[\epsilon^{1/2}] + \mathbf{Q}[\epsilon^{1/2}] \mathbf{f}[\epsilon^{1/2}]$$

- (i) The formal series of deformation patterns $\xi[\epsilon^{1/2}]$ starts with $k = 0$ (instead of degree -2 for plates).
- (ii) The first terms of the series $\mathbf{V}[\epsilon]$ are

$$\begin{aligned} \mathbf{V}^0 \xi &= \xi, \quad \mathbf{V}^{1/2} \equiv 0, \\ \mathbf{V}^1 \xi &= (-X_3(D_a \xi_3 + 2b_a^b \xi_b), P_m(X_3) \gamma_a^m(\xi)) \end{aligned} \quad (56)$$

where P_m is the polynomial defined in (9), and the tensors D (covariant derivative), b (curvature), and γ (change of metric) are introduced in Section 4.1. Even if the displacement $\mathbf{V}^1 \xi$ is given through its components in a local coordinate system, it indeed defines an intrinsic displacement, since D_a, b_a^b , and γ_a^b are well-defined independently of the choice of a local parameterization of the surface. Note that $\gamma_a^b(\xi)$ in (56) degenerates to $\text{div } \xi_T$ in the case of plates where $b_{ab} = 0$. More generally, for all integer $k \geq 0$, all 'odd' terms $\mathbf{V}^{k+1/2}$ are zero and, if $b = 0$, all even terms \mathbf{V}^k degenerate to the operators in (11). In particular, their degrees are the same as in (11).

- (iii) The formal series $\mathbf{Q}[\epsilon^{1/2}]$ appears as a generalization of (13) and $\mathbf{f}[\epsilon^{1/2}]$ is the formal Taylor expansion of \mathbf{f} around the midsurface $x_3 = 0$, which means that for all integer $k \geq 0$, $\mathbf{f}^{k+1/2} \equiv 0$ and \mathbf{f}^k is given by (12).

4.2.3 Membrane deformation patterns

The first term ξ^0 solves the membrane equation

$$\xi^0 \in H_0^1 \times H_0^1 \times L^2(S), \quad \forall \zeta' \in H_0^1 \times H_0^1 \times L^2(S), \quad a_{S,m}(\xi^0, \zeta') = 2 \int_S \xi' \cdot \mathbf{f}^0 \quad (57)$$

where $\mathbf{f}^0 = \mathbf{f}|_S$ and $a_{S,m}$ is the membrane form

$$a_{S,m}(\xi, \zeta') = 2 \int_S \tilde{\Lambda}^{\text{mem}} \gamma_{ab}(\xi) \gamma^{ab}(\zeta') dS \quad (58)$$

with the reduced energy material tensor on the midsurface (with $\tilde{\Lambda}$ still given by (17)):

$$\tilde{\Lambda}^{\text{mem}} = \tilde{\Lambda}^{\text{ab}} a^{\text{ab}} + \mu (a^{\text{ab}} a^{\text{ab}} + a^{\text{ab}} a^{\text{ba}})$$

Problem (57) can be equivalently formulated as $\mathbf{L}^0 \xi^0 = \mathbf{f}^0$ with Dirichlet boundary conditions $\xi_T^0 = 0$ on ∂S and is corresponding to the membrane equations on plates (compare with (19) and (22)). The operator \mathbf{L}^0 is called membrane operator and, thus, the change of metric $\gamma_{ab}(\xi)$ with respect to the deformation pattern ξ appears to coincide with the membrane strain tensor; see Naghdi (1963) and Koiter (1970a). If $b = 0$, the third component of $\mathbf{L}^0 \xi$ vanishes while the surface part degenerates to the membrane operator (20). In the general case, the properties of \mathbf{L}^0 depends on the geometry of S : \mathbf{L}^0 is elliptic (of multidegree (2, 2, 0) in the sense of Agmon, Douglis and Nirenberg, 1964) in \mathbf{x} if and only if S is elliptic in \mathbf{x} ; see Ciarlet (2000), Goeleven (1996), and Sanchez-Hubert and Sanchez-Palencia (1997).

As in (21), the formal series $\xi[\epsilon^{1/2}]$ solves a reduced equation on the midsurface with formal series $\mathbf{L}[\epsilon^{1/2}]$, $\mathbf{R}[\epsilon^{1/2}]$, $\mathbf{f}[\epsilon^{1/2}]$ and $\mathbf{d}[\epsilon^{1/2}]$, degenerating to the formal series (21) in the case of plates.

4.2.4 Boundary layer terms

Problem (57) cannot solve for the boundary conditions $\xi_{3|S}^0 = \partial_a \xi_{3|S}^0 = 0$ (see the first terms in (24)). The two-dimensional boundary layer terms $\mathbf{q}^{k/2}$ compensate these nonzero traces: We have for $k = 0$,

$$\begin{aligned} \mathbf{q}^0 &= (0, \mathbf{q}_3^0(T, s)) \quad \text{with} \quad \mathbf{q}_3^0(0, s) = -\xi_{3|S}^0 \\ \text{and} \quad \partial_a \mathbf{q}_3^0(0, s) &= 0 \end{aligned}$$

For $k = 1$, the trace $\partial_a \mathbf{q}_3^1|_{S}$ is compensated by $\mathbf{q}_3^{1/2}$: The scale $\epsilon^{1/2}$ arises from these surface boundary layer terms. More generally, the terms $\mathbf{q}^{k/2}$ are polynomials of degree $[k/2]$ in X_3 and satisfy

$$|\epsilon^{\mathbf{T}} \mathbf{q}(\mathbf{T}, s, X_3)| \quad \text{bounded as} \quad T \rightarrow \infty$$

for all $\eta < (3\mu(\tilde{\Lambda} + \mu))^{1/4} (\tilde{\Lambda} + 2\mu)^{-1/2} b_{\text{mem}}(0, s)^{1/2}$ where $b_{\text{mem}}(0, s) > 0$ is the tangential component of the curvature tensor along ∂S .

The three-dimensional boundary layer terms $\mathbf{w}^{k/2}$ have a structure similar to the case of plates. The first nonzero term is \mathbf{w}^1 .

4.2.5 The Koiter model

Koiter (1960) proposed the solution \mathbf{z}^ε of following surface problem

Find $\mathbf{z}^\varepsilon \in V_K(S)$ such that

$$ea_{S,m}(\mathbf{z}^\varepsilon, \mathbf{z}') + e^3 a_{S,b}(\mathbf{z}^\varepsilon, \mathbf{z}') = 2\varepsilon \int_S \mathbf{z}' \cdot \mathbf{f}^0, \quad \forall \mathbf{z}' \in V_K(S) \quad (59)$$

to be a good candidate for approximating the three-dimensional displacement by a two-dimensional one. Here the variational space is

$$V_K(S) := H_0^1 \times H_0^1 \times H_0^2(S) \quad (60)$$

and the bilinear form $a_{S,b}$ is the bending form:

$$a_{S,b}(\mathbf{z}, \mathbf{z}') = \frac{2}{3} \int_S \tilde{A}^{\text{bend}} \rho_{ab}(\mathbf{z}) \rho_{ab}(\mathbf{z}') dS \quad (61)$$

Note that the operator underlying problem (59) has the form $\mathbf{K}(\varepsilon) = \varepsilon \mathbf{L}^0 + \varepsilon^3 \mathbf{B}$ where the membrane operator \mathbf{L}^0 is the same as in (57) and the bending operator \mathbf{B} is associated with $a_{S,b}$. Thus, the change of curvature tensor ρ_{ab} appears to be identified with the bending strain tensor. Note that $\mathbf{K}(\varepsilon)$ is associated with the two-dimensional energy (compare with (44))

$$2\varepsilon \int_S \tilde{A}^{\text{bend}} \gamma_{ab}(\mathbf{z}) \gamma_{ab}(\mathbf{z}') dS + \frac{2\varepsilon^3}{3} \int_S \tilde{A}^{\text{bend}} \rho_{ab}(\mathbf{z}) \rho_{ab}(\mathbf{z}') dS \quad (62)$$

For ε small enough, the operator $\mathbf{K}(\varepsilon)$ is elliptic of multi-degree (2, 2, 4) and is associated with the Dirichlet conditions $\mathbf{z} = 0$ and $\partial_n \mathbf{z}_3 = 0$ on ∂S . The solution \mathbf{z}^ε of the Koiter model for the clamped case solves equivalently the equations

$$(\mathbf{L}^0 + \varepsilon^3 \mathbf{B}) \mathbf{z}^\varepsilon(\mathbf{x}_\tau) = \mathbf{f}^0(\mathbf{x}_\tau) \text{ on } S \text{ and } \mathbf{z}^\varepsilon|_{\partial S} = 0, \quad \partial_n \mathbf{z}_3|_{\partial S} = 0 \quad (63)$$

This solution has also a multiscale expansion given by the following theorem.

Theorem 5. (Faou, 2003) For the solutions of problem (63), $\varepsilon \in (0, \varepsilon_0]$, there exist regular terms $\mathbf{z}^{k/2}(\mathbf{x}_\tau)$ and boundary layer terms $\psi^{k/2}(\mathbf{T}, \mathbf{s})$, $k \geq 0$, such that

$$\mathbf{z}^\varepsilon \simeq \mathbf{z}^0 + \chi \psi^0 + \varepsilon^{1/2}(\mathbf{z}^{1/2} + \chi \psi^{1/2}) + \varepsilon^1(\mathbf{z}^1 + \chi \psi^1) + \dots \quad (64)$$

in the sense of asymptotic expansions: The following estimates hold

$$\left\| \mathbf{z}^\varepsilon - \sum_{k=0}^{2K} \varepsilon^{k/2}(\mathbf{z}^{k/2} + \chi \psi^{k/2}) \right\|_{\varepsilon S} \leq C_K(\mathbf{f}) \varepsilon^{K+1/4},$$

$$K = 0, 1, \dots$$

where $\|\mathbf{z}\|_{\varepsilon S}^2 = \|\gamma(\mathbf{z})\|_{L^2(S)}^2 + \varepsilon^2 \|\rho(\mathbf{z})\|_{L^2(S)}^2$ and $C_K(\mathbf{f})$ is independent of $\varepsilon \in (0, \varepsilon_0]$.

The precise comparison between the terms in the expansions (55) and (64) shows that [1] $\xi^0 = \mathbf{z}^0$, $\xi^{1/2} = \mathbf{z}^{1/2}$, $\varphi^0 = \psi^0$, $\varphi^{1/2} = \psi^{1/2}$, while ξ^1 and \mathbf{z}^1 , $\varphi_3^{1/2}$ and $\psi_3^{1/2}$ are generically different, respectively. This allows obtaining optimal estimates in various norms: Considering the scaled domain $\Omega \simeq S \times (-1, 1)$, we have

$$\|\mathbf{u}^\varepsilon - \mathbf{z}^\varepsilon\|_{H^1 \times H^1 \times L^2(\Omega)} \leq \|\mathbf{u}^\varepsilon - \xi^0\|_{H^1 \times H^1 \times L^2(\Omega)} + \|\mathbf{z}^\varepsilon - \xi^0\|_{H^1 \times H^1 \times L^2(\Omega)} \leq C\varepsilon^{1/4} \quad (65)$$

This estimate implies the convergence result of Ciarlet and Lods (1996a) and improves the estimate in Mardare (1998a). To obtain an estimate in the energy norm, we need to reconstruct a 3-D displacement from \mathbf{z}^ε : First, the Kirchhoff-like [2] displacement associated with \mathbf{z}^ε writes, cf. (56)

$$\mathbf{U}_K^{1,0} \mathbf{z}^\varepsilon = (\mathbf{z}_\alpha^\varepsilon - x_3(D_\alpha \mathbf{z}_3^\varepsilon + 2b_\alpha^\varepsilon \mathbf{z}_3^\varepsilon), \mathbf{z}_3^\varepsilon) \quad (66)$$

and next, according to Koiter (1970a), we define the reconstructed quadratic displacement [3]

$$\mathbf{U}_K^{1,2} \mathbf{z}^\varepsilon = \mathbf{U}_K^{1,0} \mathbf{z}^\varepsilon + \frac{\lambda}{\lambda + 2\mu} \left(\mathbf{0}, -x_3 \gamma_\alpha^\varepsilon(\mathbf{z}^\varepsilon) + \frac{x_3^2}{2} \rho_\alpha^\varepsilon(\mathbf{z}^\varepsilon) \right) \quad (67)$$

Then there holds (compare with (46) for plates):

$$\|\mathbf{u}^\varepsilon - \mathbf{U}_K^{1,2} \mathbf{z}^\varepsilon\|_{E(\Omega^\varepsilon)} \leq C\sqrt{\varepsilon} \|\mathbf{u}^\varepsilon\|_{E(\Omega^\varepsilon)} \quad (68)$$

and similar to plates, the error factor $\sqrt{\varepsilon}$ is optimal and is due to the first boundary layer term \mathbf{w}^1 . Moreover, expansion (64) allows proving that the classical models discussed in Budiansky and Sanders (1967), Naghdi (1963), Novozhilov (1959), and Koiter (1970a) have all the same convergence rate (68).

4.3 Convergence results for general shells

We still embed Ω^ε in the family (54) with S the mid-surface of Ω^ε . The fact that all the classical models are equivalent for clamped elliptic shells may not be true in more general cases, when the shell becomes sensitive (e.g. for a partially clamped elliptic shell with a free portion in its lateral surface) or produces bending effects (case of parabolic or hyperbolic shells with adequate lateral boundary conditions).

4.3.1 Surface membrane and bending energy

Nevertheless, the Koiter model seems to keep good approximation properties with respect to the 3-D model. The variational space V_K of the Koiter model is, in the totally clamped case given by the space $V_K(S)$ (60) (if the shell Ω^ε is clamped only on the part $\gamma_0 \times (-\varepsilon, \varepsilon)$ of its boundary (with $\gamma_0 \subset \partial S$), the Dirichlet boundary conditions in the space V_K have to be imposed only on γ_0). As already mentioned (62), the Koiter model is associated with the bilinear form $ea_{S,m} + e^3 a_{S,b}$ with $a_{S,m}$ and $a_{S,b}$ the membrane and bending forms defined for $\mathbf{z}, \mathbf{z}' \in V_K(S)$ by (58) and (61) respectively.

From the historical point of view, such a decomposition into a membrane (or stretching) energy and a bending energy on the midsurface was first derived by Love (1944) in principal curvature coordinate systems, that is, for which the curvature tensor (b_α^β) is diagonalized. The expression of the membrane energy proposed by Love is the same as (61), in contrast with the bending part for which the discussion was very controversial; see Budiansky and Sanders (1967), Novozhilov (1959), Koiter (1960), and Naghdi (1963) and the reference therein. Koiter (1960) gave the most natural expression using intrinsic tensor representations: The Koiter bending energy only depends on the change of curvature tensor ρ_{ab} , in accordance with Bonnet theorem characterizing a surface by its metric and curvature tensors a_{ab} and b_{ab} ; see Stoker (1969).

For any geometry of the midsurface S , the Koiter model in its variational form (59) has a unique solution; see Bernadou and Ciarlet (1976).

4.3.2 Classification of shells

According to this principle each shell, in the zero thickness limit, concentrates its energy either in the bending surface energy $a_{S,b}$ (flexural shells) or in the membrane surface energy $a_{S,m}$ (membrane shells).

The behavior of the shell depends on the 'inextensional displacement' space

$$V_F(S) := \{\xi \in V_K(S) \mid \gamma_{ab}(\xi) = 0\} \quad (69)$$

The key role played by this space is illustrated by the following fundamental result:

Theorem 6.

- (i) (Sanchez-Hubert and Sanchez-Palencia, 1997; Ciarlet, Lods and Miara, 1996) Let \mathbf{u}^ε be the solution of problem (3). In the scaled domain $\Omega \simeq S \times (-1, 1)$, the displacement $\varepsilon^2 \mathbf{u}^\varepsilon(\mathbf{x}_\tau, \mathbf{x}_3)$ converges in $H^1(\Omega)^3$ as $\varepsilon \rightarrow 0$. Its limit is given by the solution $\xi^{-2} \in V_F(S)$

of the bending problem

$$\forall \xi' \in V_F(S) \quad a_{S,b}(\xi^{-2}, \xi') = 2 \int_S \xi' \cdot \mathbf{f}^0 \quad (70)$$

- (ii) (Ciarlet and Lods, 1996b) Let \mathbf{z}^ε be the solution of problem (59). Then $\varepsilon^2 \mathbf{z}^\varepsilon$ converges to ξ^{-2} in $V_K(S)$ as $\varepsilon \rightarrow 0$.

A shell is said *flexural* (or *noninhibited*) when $V_F(S)$ is not reduced to $\{0\}$. Examples are provided by cylindrical shells (or portions of cones) clamped along their generatrices and free elsewhere. Of course, plates are flexural shells according to the above definition since in that case, $V_F(S)$ is given by $\{\xi = (0, \xi_3) \mid \xi_3 \in H_0^2(S)\}$ and the bending operator (70) coincides with the operator (16).

In the case of clamped elliptic shells, we have $V_F(S) = \{0\}$. For these shells, \mathbf{u}^ε and \mathbf{z}^ε converge in $H^1 \times H^1 \times L^2$ to the solution ξ^0 of the membrane equation (57); see Ciarlet and Lods (1996a) and (65): Such shells are called *membrane shells*. The other shells for which $V_F(S)$ reduces to $\{0\}$ are called *generalized membrane shells* (or *inhibited shells*) and for these also, a delicate functional analysis provides convergence results to a membrane solution in spaces with special norms depending on the geometry of the midsurface; see Ciarlet and Lods (1996a) and Ciarlet (2000), Ch. 5. It is also proved that the Koiter model converges in the same sense to the same limits; see Ciarlet (2000), Ch. 7.

Thus, plates and elliptic shells represent extreme situations: Plates are a pure bending structures with an inextensional displacement space as large as possible, while clamped elliptic shells represent a pure membrane situation where $V_F(S)$ reduces to $\{0\}$ and where the membrane operator is elliptic.

4.4 Shallow shells

We make a distinction between 'physical' shallow shells in the sense of Ciarlet and Paumier (1986) and 'mathematical' shallow shells in the sense of Pitkäranta, Mäkelä and Schwab (2001). The former involves shells with a curvature tensor of the same order as the thickness, whereas the latter addresses a boundary value problem obtained by freezing coefficients of the Koiter problem at one point of a standard shell.

4.4.1 Physical shallow shells

Let R denote the smallest principal radius of curvature of the midsurface S and let D denote the diameter of S . As proved in Andreoli and Faou (2001) if there holds

$$R \geq 2 \cdot D \quad (71)$$

then there exists a point $\mathbf{p}_\tau \in S$ such that the orthogonal projection of S on its tangent plan in \mathbf{p}_τ allows the representation of S as a C^∞ graph in \mathbb{R}^3 :

$$\omega \ni (x_1, x_2) \mapsto (x_1, x_2, \Theta(x_1, x_2)) := \mathbf{x}_\tau \in S \subset \mathbb{R}^3 \quad (72)$$

where ω is an immersed (in particular, ω may have self-intersection) domain of the tangent plane in \mathbf{p}_τ , and where Θ is a function on this surface. Moreover, we have

$$|\Theta| \leq CR^{-1} \quad \text{and} \quad \|\nabla \Theta\| \leq CR^{-1} \quad (73)$$

with constants C depending only on D .

We say that Ω^ε is a *shallow shell* if S satisfies a condition of the type

$$R^{-1} \leq Cd \quad (74)$$

where C does not depend on d . Thus, if S is a surface satisfying (74), for d sufficiently small S satisfies (71) whence representation (72). Moreover, (73) yields that Θ and $\nabla \Theta$ are $\lesssim d$. In these conditions, we can choose to embed Ω^ε into another family of thin domains than (54): We set $\theta = d^{-1}\Theta$ and define for any $\varepsilon \in (0, d]$ the surface S^ε by its parameterization (cf. (72))

$$\omega \ni (x_1, x_2) \mapsto (x_1, x_2, \varepsilon \theta(x_1, x_2)) := \mathbf{x}_\tau \in S^\varepsilon$$

It is natural to consider Ω^ε as an element of the family Ω^ε given as the image of the application

$$\omega \times (-\varepsilon, \varepsilon) \ni (x_1, x_2, x_3) \mapsto (x_1, x_2, \varepsilon \theta(x_1, x_2) + x_3 \mathbf{n}^\varepsilon(\mathbf{x}_\tau)) \quad (75)$$

where $\mathbf{n}^\varepsilon(\mathbf{x}_\tau)$ denotes the unit normal vector to the mid-surface S^ε . We are now in the framework of Ciarlet and Pammier (1986).

A multiscale expansion for the solution of (3) is given in Andreoli and Faou (2001). The expansion is close to that of plates, except that the membrane and bending operators yielding the deformation patterns are linked by lower-order terms: The associated membrane and bending strain components $\tilde{\gamma}_{\text{eb}}$ and $\tilde{\rho}_{\text{eb}}$ are respectively given by

$$\tilde{\gamma}_{\text{eb}} := \frac{1}{2}(\partial_1 z_\theta + \partial_2 z_\theta) - \varepsilon \partial_\theta \theta z_3 \quad \text{and} \quad \tilde{\rho}_{\text{eb}} := \partial_\theta z_3 \quad (76)$$

It is worth noticing that the above strains are asymptotic approximations of the Koiter membrane and bending strains associated with the midsurface $S = S^\varepsilon$. As a consequence, the Koiter model and the three-dimensional equations converge to the same Kirchhoff–Love limit.

4.4.2 Mathematical shallow shells

These models consist in freezing coefficients of standard two-dimensional models at a given point $\mathbf{p}_\tau \in S$ in a principal curvature coordinate system. That procedure yields, with $b_i := \kappa_i(\mathbf{p}_\tau)$:

$$\begin{aligned} \gamma_{11} &= \partial_1 z_1 - b_1 z_3, & \gamma_{22} &= \partial_2 z_2 - b_2 z_3, \\ \gamma_{12} &= \frac{1}{2}(\partial_1 z_2 + \partial_2 z_1) \end{aligned} \quad (77)$$

for the membrane strain tensor, and

$$\begin{aligned} \kappa_{11} &= \partial_1^2 z_3 + b_1 \partial_1 z_3, & \kappa_{22} &= \partial_2^2 z_3 + b_2 \partial_2 z_3, \\ \kappa_{12} &= \partial_1 \partial_2 z_3 + b_1 \partial_2 z_1 + b_2 \partial_1 z_2 \end{aligned} \quad (78)$$

as a simplified version of the bending strain tensor. Such a localization procedure is considered as valid if the diameter D is small compared to R

$$R \gg D \quad (79)$$

and for the case of cylindrical shells where the strains have already the form (77)–(78) in cylindrical coordinates (see equation (80) below). In contrast with the previous one, this notion of shallowness does not refer to the thickness. Here R is not small, but D is. Such objects are definitively shells and are not plate-like.

These simplified models are valuable so to develop numerical approximation methods, (Havu and Pitkäranta, 2002, 2003) and to find possible boundary layer length scales, (Pitkäranta, Matsche and Schwab, 2001): These length scales (width of transition regions from the boundary into the interior) at a point $\mathbf{p}_\tau \in \partial S$ are $\varepsilon^{1/2}$ in the non-degenerate case ($b_{\text{as}}(\mathbf{p}_\tau) \neq 0$), $\varepsilon^{1/3}$ for hyperbolic degeneration (\mathbf{p}_τ hyperbolic and $b_{\text{as}}(\mathbf{p}_\tau) = 0$) and $\varepsilon^{1/4}$ for parabolic degeneration (\mathbf{p}_τ parabolic and $b_{\text{as}}(\mathbf{p}_\tau) = 0$).

To compare with the standard shell equations, note that in the case of an axisymmetric shell whose midsurface is represented by

$$\Psi: (x_1, x_2) \mapsto (f(x_1) \cos x_2, f(x_1) \sin x_2, x_1)$$

where $x_1 \in \mathbb{R}$, $x_2 \in [0, 2\pi]$ and $f(x_1) > 0$ is a smooth function, we have

$$\begin{aligned} \gamma_{11}(\mathbf{x}) &= \partial_1 z_1 - \frac{f'(x_1)f''(x_1)}{1+f'(x_1)^2} z_1 + \frac{f''(x_1)}{\sqrt{1+f'(x_1)^2}} z_3 \\ \gamma_{22}(\mathbf{x}) &= \partial_2 z_2 + \frac{f(x_1)f'(x_1)}{1+f'(x_1)^2} z_1 - \frac{f'(x_1)}{\sqrt{1+f'(x_1)^2}} z_3 \\ \gamma_{12}(\mathbf{x}) &= \frac{1}{2}(\partial_1 z_2 + \partial_2 z_1) - \frac{f'(x_1)}{f(x_1)} z_2 \end{aligned} \quad (80)$$

Hence the equation (77) is exact for the case of cylindrical shells, where $f(x_1) \equiv R > 0$, and we can show that the same holds for (78).

4.5 Versatility of Koiter model

On any midsurface S , the deformation pattern \mathbf{z}^ε solution of the Koiter model (59) exists. In general, the mean value of the displacement \mathbf{u}^ε through the thickness converges to the same limit as \mathbf{z}^ε when $\varepsilon \rightarrow 0$ in a weak sense depending on the type of the midsurface and the boundary conditions; see Ciarlet (2000). Nevertheless, actual convergence results hold in energy norm when considering reconstructed displacement from the deformation pattern \mathbf{z}^ε .

4.5.1 Convergence of the Koiter reconstructed displacement

On any midsurface S , the three-dimensional Koiter reconstructed displacement $\mathbf{U}_K^{1,2,\varepsilon} \mathbf{z}^\varepsilon$ is well-defined by (66)–(67). Let us set

$$\mathbf{e}(S, \varepsilon, \mathbf{z}^\varepsilon, \mathbf{u}^\varepsilon) := \frac{\|\mathbf{u}^\varepsilon - \mathbf{U}_K^{1,2,\varepsilon} \mathbf{z}^\varepsilon\|_{E(\Omega^\varepsilon)}}{\|\mathbf{z}^\varepsilon\|_{E(S)}} \quad (81)$$

with $\|\mathbf{z}\|_{E(S)}$, the square root of the Koiter energy (62). In Koiter (1970a,b), an estimate is given: $\mathbf{e}(S, \varepsilon, \mathbf{z}^\varepsilon, \mathbf{u}^\varepsilon)^2$ would be bounded by $\varepsilon R^{-1} + \varepsilon^2 L^{-2}$, with R the smallest principal radius of curvature of S , and L the smallest wavelength of \mathbf{z}^ε . It turns out that in the case of plates, we have $L = O(1)$, $R^{-1} = 0$ and, since (46) is optimal, the estimate fails. In contrast, in the case of clamped elliptic shells, we have $L = O(\sqrt{\varepsilon})$, $R^{-1} = O(1)$ and the estimate gives back (68).

Two years after the publications of Koiter (1970a,b), it was already known that the above estimate does not hold as $\varepsilon \rightarrow 0$ for plates. We read in Koiter and Simmonds (1973) ‘The somewhat depressing conclusion for most shell problems is, similar to the earlier conclusions of GOLDENVEIZER, that no better accuracy of the solutions can be expected than of order $\varepsilon L^{-1} + \varepsilon R^{-1}$, even if the equations of first-approximation shell theory would permit, in principle, an accuracy of order $\varepsilon^2 L^{-2} + \varepsilon R^{-1}$ ’.

The reason for this is also explained by John (1971) in these terms: ‘Concentrating on the interior we sidestep all kinds of delicate questions, with an attendant gain in certainty and generality. The information about the interior behavior can be obtained much more cheaply (in the mathematical sense) than that required for the discussion of boundary value problems, which form a more ‘transcendental’ stage’.

Koiter’s tentative estimate comes from formal computations also investigated by John (1971). The analysis by operator formal series introduced in Faou (2002) is in the same spirit: For any geometry of the midsurface, there exist formal series $V[\varepsilon]$, $\mathbf{H}[\varepsilon]$, $\mathbf{Q}[\varepsilon]$, and $\mathbf{L}[\varepsilon]$ reducing the three-dimensional formal series problem to a two-dimensional problem of the form (21) with $\mathbf{L}[\varepsilon] = \mathbf{L}^0 + \varepsilon^2 \mathbf{L}^2 + \dots$ where \mathbf{L}^0 is the membrane operator associated with the form (58). The bending operator \mathbf{B} associated with $a_{S,0}$ can be compared to the operator \mathbf{L}^2 appearing in the formal series $\mathbf{L}[\varepsilon]$: We have

$$\forall \zeta, \zeta' \in V_F(S) \quad (\mathbf{L}^2 \zeta, \zeta')_{L^2(S)} = (\mathbf{B} \zeta, \zeta')_{L^2(S)} \quad (82)$$

Using this formal series analysis, the first two authors are working on the derivation of a sharp expression of $\mathbf{e}(S, \varepsilon, \mathbf{z}^\varepsilon, \mathbf{u}^\varepsilon)$ including boundary layers effects, and optimal in the case of plates and clamped elliptic shells; see Dauge and Faou (2004).

In this direction also, Lods and Mardare (2002) prove the following estimate for totally clamped shells

$$\|\mathbf{u}^\varepsilon - (\mathbf{U}_K^{1,2,\varepsilon} \mathbf{z}^\varepsilon + \mathbf{w}^\varepsilon)\|_{E(\Omega^\varepsilon)} \leq C \varepsilon^{1/4} \|\mathbf{u}^\varepsilon\|_{E(\Omega^\varepsilon)} \quad (83)$$

with \mathbf{w}^ε an explicit boundary corrector of $\mathbf{U}_K^{1,2,\varepsilon} \mathbf{z}^\varepsilon$.

4.5.2 Convergence of Koiter eigenvalues

The operator $\varepsilon^{-1} \mathbf{K}(\varepsilon)$ has a compact inverse, therefore its spectrum is discrete with only an accumulation point at $+\infty$. We agree to call Koiter eigenvalues the eigenvalues of the former operator, that is, the solutions μ^ε of

$$\exists \mathbf{z}^\varepsilon \in V_K(S) \setminus \{0\} \text{ such that}$$

$$a_{S,0}(\mathbf{z}^\varepsilon, \mathbf{z}^\varepsilon) + \varepsilon^2 a_{S,2}(\mathbf{z}^\varepsilon, \mathbf{z}^\varepsilon) = 2\mu^\varepsilon \int_S \mathbf{z}^\varepsilon \cdot \mathbf{z}^\varepsilon, \quad \forall \mathbf{z}^\varepsilon \in V_K(S) \quad (84)$$

As already mentioned in Section 2.4, cf. (31), this spectrum provides the limiting behavior of three-dimensional eigenvalues for plates. Apparently, very little is known for general shells.

Concerning Koiter eigenvalues, interesting results are provided by Sanchez-Hubert and Sanchez-Palencia (1997), Ch. X: The μ^ε are attracted by the spectrum of the membrane operator $\mathbf{S}(\mathbf{M})$ where \mathbf{M} is the self-adjoint unbounded operator associated with the symmetric bilinear form $a_{S,0}$ defined on the space $H^1 \times H^1 \times L^2(S)$. There holds (we still assume that S is smooth up to its boundary):

Theorem 7. *The operator $\mathbf{M} + \mu \text{Id}$ is elliptic of multidegree $(2, 2, 0)$ for $\mu > 0$ large enough. Moreover its essential spectrum $\mathbf{S}_\infty(\mathbf{M})$ satisfies:*

- (i) If S is elliptic and clamped on its whole boundary, $\mathfrak{E}_a(\mathbf{M})$ is a closed interval $[a, b]$, with $a > 0$.
- (ii) If S is elliptic and not clamped on its whole boundary, $\mathfrak{E}_a(\mathbf{M}) = \{0\} \cup [a, b]$ with $a > 0$.
- (iii) For any other type of shell (i.e. there exists at least one point where the Gaussian curvature is ≤ 0) $\mathfrak{E}_a(\mathbf{M})$ is a closed interval of the form $[0, b]$, with $b \geq 0$.

If the shell is of flexural type, the lowest eigenvalues μ_ε tend to 0 like $O(\varepsilon^2)$, same as for plates; see (29). If the shell is clamped elliptic, the μ_ε are bounded from below by a positive constant independent of ε . In any other situation, we expect that the lowest μ_ε still tends to 0 as $\varepsilon \rightarrow 0$.

5 HIERARCHICAL MODELS FOR SHELLS

The idea of deriving hierarchical models for shells goes back to Vekua (1955, 1965, 1985) and corresponds to classical techniques in mechanics: Try to find asymptotic expansions in x_3 by use of Taylor expansion around the midsurface S . An alternative approach consists in choosing the coefficients z_j^ε in (38) as moments through the thickness against Legendre polynomials $L_n(x_3/\varepsilon)$.

Vogelius and Babuška (1981a, 1981b, 1981c) laid the foundations of hierarchical models in view of their applications to numerical analysis (for scalar problems).

5.1 Hierarchies of semidiscrete subspaces

The concepts mentioned in Section 3 can be adapted to the case of shells. In contrast with plates for which there exist convenient Cartesian system of coordinates fitting with the tangential and normal directions to the midsurface, more nonequivalent options are left open for shells. They are; for example;

The direction of semidiscretization: The intrinsic choice is of course the normal direction to the midsurface (variable x_3), nevertheless for shells represented by a single local chart like in (72), any transverse direction could be chosen. In the sequel, we only consider semidiscretizations in the normal direction.

The presence or absence of privileged components for the displacement field in the Ansatz (38). If one privileged component is chosen, it is of course the normal one u_3 and the two other ones are $(u_\alpha) = u_\tau$. Then the sequence of orders \mathbf{q} is of the form $\mathbf{q} = (q_\tau, q_\tau, q_3)$, and the space $V^{\mathbf{q}}(\Omega^\varepsilon)$ has the form (48). Note that this space is independent of the choice of local coordinates on S .

If there is no privileged component, \mathbf{q} has to be of the form (q, q, q) and the space $V^{\mathbf{q}}(\Omega^\varepsilon)$ can be written

$$V^{\mathbf{q}}(\Omega^\varepsilon) = \left\{ \mathbf{u} = (u_1, u_2, u_3) \in V(\Omega^\varepsilon), \right. \\ \exists \mathbf{x}^n = (x_1^n, x_2^n, x_3^n) \in H_0^1(S)^3, \quad 0 \leq n \leq q, \\ \left. u_j = \sum_{n=0}^q x_3^n z_j^n(x_\tau), \quad j = 1, 2, 3 \right\} \quad (85)$$

Here, for ease of use, we take Cartesian coordinates, but the above definition is independent of any choice of coordinates in $\Omega^\varepsilon \subset \mathbb{R}^3$. In particular, it coincides with the space (48) for $q_\tau = q_3$.

Then the requirements of approximability (34), asymptotic consistency (35), and optimality of the convergence rate (36) make sense.

5.2 Approximability

For any fixed thickness ε , the approximability issue is as in the case of plates. By Céa's lemma, there exists an adimensional constant $C > 0$ depending only on the Poisson ratio ν , such that

$$\|\mathbf{u}^\varepsilon - \mathbf{u}^{\mathbf{q},\varepsilon}\|_{E(\Omega^\varepsilon)} \leq C \|\mathbf{u}^\varepsilon - \mathbf{v}^{\mathbf{q}}\|_{E(\Omega^\varepsilon)} \quad \forall \mathbf{v}^{\mathbf{q}} \in V^{\mathbf{q}}(\Omega^\varepsilon)$$

and the determination of approximability properties relies on the construction of a best approximation of \mathbf{u}^ε by functions in $V^{\mathbf{q}}(\Omega^\varepsilon)$.

In Avalishvili and Gordeziani (2003), approximability is proved using the density of the sequence of spaces $V^{\mathbf{q}}(\Omega^\varepsilon)$ in $H^1(\Omega^\varepsilon)^3$. But the problem of finding a rate for the convergence in (33) is more difficult, since the solution \mathbf{u}^ε has singularities near the edges and, consequently, does not belong to $H^2(\Omega^\varepsilon)$ in general. For scalar problems, Vogelius and Babuška (1981a, 1981b, 1981c) prove best approximation results using weighted Sobolev norms [4]. Up to now, for elasticity systems, there are no such results taking the actual regularity of \mathbf{u}^ε into account.

It is worth noticing that, in order to obtain an equality of the form (36), we must use Korn inequality, since most approximation results are based on Sobolev norms. But due to blow up of the Korn constant when $\varepsilon \rightarrow 0$, it seems hard to obtain sharp estimates in the general case. (Let us recall that it behaves as ε^{-1} in the case of partially clamped shells.)

5.3 Asymptotic consistency

Like for plates, the presence of the nonpolynomial three-dimensional boundary layers \mathbf{w}^ε generically produces a

limitation in the convergence rate in (35). As previously mentioned, the only case where a sharp estimate is available, is the case of clamped elliptic shells. Using (68), we indeed obtain the following result (compare with Theorem 3):

Theorem 8. *If the midsurface S is elliptic, if the shell is clamped along its whole lateral boundary, and if $\mathbf{f}|_S \equiv 0$, then for any $\mathbf{q} \geq (1, 1, 2)$ with definition (48), and for any $\mathbf{q} \geq (2, 2, 2)$ with (85), and with the use of the standard 3-D elastic energy (2), there exists $C_q = C_q(\mathbf{f}) > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$*

$$\|\mathbf{u}^\varepsilon - \mathbf{u}^{\mathbf{q},\varepsilon}\|_{E(\Omega^\varepsilon)} \leq C_q \sqrt{\varepsilon} \|\mathbf{u}^\varepsilon\|_{E(\Omega^\varepsilon)} \quad (86)$$

Note that in this case, the two-dimensional boundary layers are polynomial in x_3 . Therefore they can be captured by the semidiscrete hierarchy of spaces $V^{\mathbf{q}}$.

Using estimate (83) of Lods and Mardare (2002), together with the fact that the corrector term \mathbf{w}^ε is polynomial in x_3 of degree $(0, 0, 2)$, we obtain a proof for the asymptotic consistency for any (smooth) clamped shell without assuming that the midsurface is elliptic:

Theorem 9. *If the shell is clamped along its whole lateral boundary, and if $\mathbf{f}|_S \equiv 0$, then for \mathbf{q} as in Theorem 8 and the standard energy (2), there exists $C_q = C_q(\mathbf{f}) > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$*

$$\|\mathbf{u}^\varepsilon - \mathbf{u}^{\mathbf{q},\varepsilon}\|_{E(\Omega^\varepsilon)} \leq C_q \varepsilon^{1/4} \|\mathbf{u}^\varepsilon\|_{E(\Omega^\varepsilon)} \quad (87)$$

5.4 Examples of hierarchical models

Various models of degree $(1, 1, 0)$, $(1, 1, 2)$, and $(2, 2, 2)$ are introduced and investigated in the literature. Note that the model $(1, 1, 1)$ is strictly forbidden for shells because it cannot be associated with any correct energy; see Chapelle, Ferent and Bathe (2004).

5.4.1 (1, 1, 0) models

One of the counterparts of Reissner-Mindlin model for plates is given by the Naghdi model: see Naghdi (1963, 1972). The space of admissible displacements is

$$V^{\mathbf{q}}(\Omega^\varepsilon) = \{ \mathbf{u} \in V(\Omega^\varepsilon), \exists \mathbf{z} \in H_0^1(S)^3, \exists \theta_\alpha \in H_0^1(S)^2, \\ \mathbf{u} = (\mathbf{z}_\alpha - x_3(\theta_\alpha + b_\alpha^\beta x_\beta), z_3) \} \quad (88)$$

As in (47) the energy splits into three parts (with the shear correction factor κ):

$$\begin{aligned} \tilde{\mathcal{U}}(\mathbf{u}, \mathbf{u}) = & 2\varepsilon \int_S \tilde{A}^{\text{mem}} \gamma_{\alpha\beta}(\mathbf{z}_\tau) \gamma_{\alpha\beta}(\mathbf{z}_\tau) dS \\ & + \varepsilon \kappa \int_S \mu \alpha^{\alpha\beta} (D_\alpha z_3 + b_\alpha^\beta z_\beta - \theta_\alpha) \\ & \times (D_\beta z_3 + b_\beta^\gamma z_\gamma - \theta_\beta) dS \\ & + \frac{2\varepsilon^2}{3} \int_S \tilde{A}^{\text{bend}} \bar{\rho}_{\alpha\beta}(\mathbf{z}, \theta) \bar{\rho}_{\alpha\beta}(\mathbf{z}, \theta) dS \end{aligned} \quad (89)$$

(membrane energy)
(shear energy)
(bending energy)

where

$$\bar{\rho}_{\alpha\beta}(\mathbf{z}, \theta) = \frac{1}{2} (D_\alpha \theta_\beta + D_\beta \theta_\alpha) - c_{\alpha\beta} z_3 + \frac{1}{2} b_\alpha^\gamma D_\beta z_\gamma + \frac{1}{2} b_\beta^\gamma D_\alpha z_\gamma$$

Note that when the penalization term in the shear energy goes to zero, we get $\theta_\alpha = D_\alpha z_3 + b_\alpha^\beta z_\beta$ and the displacement \mathbf{u} in (88) coincides with (66). In Lods and Mardare (2002), an estimate of the error between the solution of the Naghdi model and the solution of the 3-D model is provided in a subenergetic norm.

A more recent $(1, 1, 0)$ model (called *general shell element*; see Chapelle and Bathe, 2000) consists of the reduced energy projection on the space $V^{(1,1,0)}(\Omega^\varepsilon)$. Indeed, it does not coincide with the Naghdi model but both models possess similar asymptotic properties and they are preferred to Koster's for discretization.

5.4.2 Quadratic kinematics

In accordance with Theorems 8 and 9, it is relevant to use the standard 3-D elastic energy (2) for such kinematics. Quadratic models based on the $(1, 1, 2)$ model are investigated in Bischoff and Ramm (2000). The enrichment of the general shell element by the introduction of quadratic terms – model $(2, 2, 2)$ – is thoroughly studied from both asymptotic and numerical point views in Chapelle, Ferent and Bathe (2004) and Chapelle, Ferent and Le Tallec (2003).

6 FINITE ELEMENT METHODS IN THIN DOMAINS

We herein address some of the characteristics of finite element methods (FEM), mainly the p -version of the FEM, when applied to the primal weak formulations (3) and (33) for the solution of plate and shell models. We only address isotropic materials, although our analysis could be extended to laminated composites.

As illustrative examples, we present the results of some computations performed with the p -version FE computer program StressCheck. (StressCheck is a trade mark of Engineering Software Research and Development, Inc., 10845 Olive Blvd., Suite 170, St. Louis, MO 63141, USA.)

6.1 FEM discretizations

Let us recall that, when conformal, the FEM is a Galerkin projection into finite dimensional subspaces V_n of the variational space associated with the models under consideration. In the p -version of the FEM, subspaces are based on one partition of the domain into a finite number of subdomains $K \in \mathcal{T}$ (the mesh) in which the unknown displacement is discretized by mapped polynomial functions of increasing degree p . The subdomains K are mapped from reference element(s) \hat{K} .

6.1.1 Meshes

All finite element discretizations we consider here are based on a mesh \mathcal{T}_S of the midsurface S . We mean that the 3-D mesh of Ω^ε has in normal coordinates $(\mathbf{x}_n, \mathbf{x}_s)$ the tensor product form $[S] \times \mathcal{T}_s$ where \mathcal{T}_s represents a partition of the interval $(-\varepsilon, \varepsilon)$ in layers, for example, the two halves $(-\varepsilon, 0)$ and $(0, \varepsilon)$, or – this case is important in the sequel – , the trivial partition by only one element through the thickness. We agree to call that latter mesh a *thin element mesh*.

The 3-D elements K are thus images by maps ψ_K from reference elements \hat{K} , which are either pentahedral (triangle \times interval) or hexahedral:

$$\psi_K: \hat{K} = \hat{T} \times [0, 1] \ni (\hat{x}_1, \hat{x}_2, \hat{x}_3) \mapsto \mathbf{x} \in K$$

with \hat{T} the reference triangle or the reference square. For the 2-D FEM, we denote by T the elements in \mathcal{T}_S . They are the image of \hat{T} by maps ψ_T

$$\psi_T: \hat{T} \ni (\hat{x}_1, \hat{x}_2) \mapsto \mathbf{x}_T \in T$$

If Ω_ε is a plate, the midsurface S is plane but its boundary ∂S is not straight. For some lateral boundary conditions, for example, the hard, simple supported plate, the approximation of ∂S by a polygonal lines produces, in general, *wrong results*. This effect is known as the *Babuška paradox* (Babuška and Pitkäranta, 1990). If Ω_ε is a shell, the geometric approximation of S by 'plane' elements is also an issue: If the mappings are affine, the shell is approximated by a faceted surface which has quite different rigidity properties than the smooth surface; see Akian and Sanchez-Palencia (1992) and Chapelle and Bathe (2003), Section 6.2.

As a conclusion, good mappings have to be used for the design of the elements K (high degree polynomials or other analytic functions).

6.1.2 Polynomial spaces for hierarchical models

For hierarchical models (33), the discretization is indeed two-dimensional: The degree \mathbf{q} of the hierarchy being fixed, the unknowns of (33) are the functions \mathbf{z}_j^q defined on S and representing the displacement according to (38), where the director functions Φ_j^q form adequate bases of polynomials in one variable, for example, Legendre polynomials L_n . We have already mentioned in Section 5 that the only *intrinsic* option for the choice of components is taking $j = (\alpha, 3)$, which results into the Ansatz (written here with Legendre polynomials)

$$\mathbf{u}_T = \sum_{n=0}^{q_T} \mathbf{z}_T^q(\mathbf{x}_T) L_n\left(\frac{x_3}{\varepsilon}\right)$$

$$\mathbf{u}_3 = \sum_{n=0}^{q_3} \mathbf{z}_3^q(\mathbf{x}_T) L_n\left(\frac{x_3}{\varepsilon}\right)$$

Now the discretization consists in requiring that $\mathbf{z}_j^q|_T \circ \psi_T$, $\alpha = 1, 2$, and $\mathbf{z}_3^q|_T \circ \psi_T$ belong to the space $\mathbb{P}_p(\hat{T})$ for some p where $\mathbb{P}_p(\hat{T})$ is the space of polynomials in two variables

- of degree $\leq p$ if \hat{T} is the reference triangle,
- of partial degree $\leq p$ if \hat{T} is the reference square $[0, 1] \times [0, 1]$.

It makes sense to fix different degrees p_j in relation with $j = \alpha, 3$, and we set $\mathbf{p} = (p_1, p_2, p_3)$. When plugged back into formula (38), this discretization of the \mathbf{z}_j^q , $j = \alpha, 3$, yields a finite dimensional subspace $V_p^q(\Omega^\varepsilon)$ of $V^q(\Omega^\varepsilon)$. As already mentioned for the transverse degrees \mathbf{q} , cf. (48) and Section 5, we have to assume for coherence that $p_1 = p_2$ for shells. In the situation of plates, if T is affinely mapped from the reference square, the $\mathbf{z}_j^q|_T$ are simply given by

$$\mathbf{z}_1^q(\mathbf{x}_T) = \sum_{i,k=0}^{p_1} z_{1,ik}^q P_i(x_1) P_k(x_2)$$

$$\mathbf{z}_2^q(\mathbf{x}_T) = \sum_{i,k=0}^{p_2} z_{2,ik}^q P_i(x_1) P_k(x_2)$$

$$\mathbf{z}_3^q(\mathbf{x}_T) = \sum_{i,k=0}^{p_3} z_{3,ik}^q P_i(x_1) P_k(x_2)$$

where the $z_{j,ik}^q$ are real coefficients and P_i denotes a polynomial of degree i which is obtained from Legendre polynomials; see for example Szabó and Babuška (1991).

The discretization of hierarchical models (33) can also be done through the h -version or the h - p versions of FEM.

6.1.3 Polynomial spaces for 3-D discretization. Case of thin elements

In 3-D, on the reference element $\hat{K} = \hat{T} \times [0, 1]$, we can consider any of the polynomial spaces $\mathbb{P}_{p,q}(\hat{K}) = \mathbb{P}_p(\hat{T}) \otimes \mathbb{P}_q([0, 1])$, $p, q \in \mathbb{N}$. For the discretization of (3), each Cartesian component u_i of the displacement is sought for in the space of functions $v \in H^1(\Omega^\varepsilon)$ such that for any K in the mesh, $v|_K \circ \psi_K$ belongs to $\mathbb{P}_{p,q}(\hat{K})$. We denote by $V_{p,q}(\Omega^\varepsilon)$ the corresponding space of admissible displacements over Ω^ε .

In the situation where we have only one layer of elements over Ω^ε in the thickness (thin element mesh) with a (p, q) discretization, let us set $\mathbf{q} = (q, q, q)$ and $\mathbf{p} = (p, p, p)$. Then it is easy to see that, in the framework of semidiscrete spaces (85), we have the equality between discrete spaces:

$$V_{p,q}(\Omega^\varepsilon) = V_p^q(\Omega^\varepsilon) \quad (90)$$

In other words, thin elements are equivalent to the discretization of underlying hierarchical models. Let us insist on the following fact: For a true shell, the correspondence between the Cartesian components u_i and the tangential and transverse components (\mathbf{u}_T, u_3) is nonaffine. As a consequence, equality (90) holds only if the space $V_p^q(\Omega^\varepsilon)$ corresponds to the discretization of a hierarchical model in Cartesian coordinates.

Conversely, hierarchical models of the type $\mathbf{q} = (q, q, q)$ with the 'Cartesian' unknowns \mathbf{z}_j^q , $n = 0, \dots, q$, $j = 1, 2, 3$ can be discretized directly on S , or inherit a 3-D discretization; see Chapelle, Ferent and Bathe (2004). Numerical evidence that the p -version with anisotropic Ansatz spaces allows the analysis of three-dimensional shells with high accuracy was firstly presented in Düster, Bröker and Rank (2001).

6.1.4 FEM variational formulations

Let us fix the transverse degree \mathbf{q} of the hierarchical model. Its solution $\mathbf{u}^{\varepsilon,q}$ solves problem (33). For each $\varepsilon > 0$ and each polynomial degree \mathbf{p} , (33) is discretized by its finite dimensional subspace $V_p^q(\Omega^\varepsilon)$. Let $\mathbf{u}_p^{\varepsilon,q}$ be the solution of

$$\text{Find } \mathbf{u}_p^{\varepsilon,q} \in V_p^q(\Omega^\varepsilon) \text{ such that}$$

$$a^{\varepsilon,q}(\mathbf{u}_p^{\varepsilon,q}, \mathbf{u}') = \int_{\Omega^\varepsilon} \mathbf{f} \cdot \mathbf{u}' \, d\mathbf{x}, \quad \forall \mathbf{u}' \in V_p^q(\Omega^\varepsilon) \quad (91)$$

We can say that (91) is a sort of 3-D discretization of (33). But, indeed, the actual unknowns of (91) are the \mathbf{z}_j^q , $n = 0, \dots, q_T$, and \mathbf{z}_3^q , $n = 0, \dots, q_3$, or the \mathbf{z}_j^q for $n = 0, \dots, q$ and $j = 1, 2, 3$. Thus, (91) can be alternatively formulated as a 2-D problem involving spaces $Z_p^q(S)$ independent of ε , and a coercive bilinear form $a_p^q(\varepsilon)$ polynomial in ε .

Examples are provided by the Reissner–Mindlin model, cf. (47), the Koiter model (84), and the Naghdi model, cf. (89). The variational formulation now takes the form

$$\text{Find } \mathbf{Z} =: (\mathbf{z}^j)_{0 \leq n \leq q_j} \in Z_p^q(S) \text{ such that}$$

$$a_p^q(\varepsilon)(\mathbf{Z}, \mathbf{Z}') = F(\varepsilon)(\mathbf{f}, \mathbf{Z}'), \quad \forall \mathbf{Z}' \in Z_p^q(S) \quad (92)$$

where $F(\varepsilon)(\mathbf{f}, \mathbf{Z}')$ is the suitable bilinear form coupling loadings and test functions. Let us denote by \mathbf{Z}_p^q the solution of (92).

6.2 Locking issues

In the framework of the family of discretizations considered above, the *locking* effect is said to appear when a deterioration in the resulting approximation of $\mathbf{u}^{\varepsilon,q}$ by $\mathbf{u}_p^{\varepsilon,q}$, $\mathbf{p} \rightarrow \infty$ tends to ∞ , occurs as $\varepsilon \rightarrow 0$. Of course, a similar effect is reported in the h -version of FEM: The deterioration of the h -approximation also occurs when the thickness ε approaches zero.

Precise definition of locking may be found in Babuška and Suri (1992): It involves the locking parameter (the thickness ε in the case of plates), the sequence of finite element spaces V_p^q that comprise the extension procedure (the p -version in our case, but h and h - p versions can also be considered), and the norm in which error is to be measured. Of course, in different error measures different locking phenomena are expected.

6.2.1 Introduction to membrane locking

A locking-free approximation scheme is said to be *robust*. For a bilinear form $a_\varepsilon(\varepsilon)$ of the form $a_0 + \varepsilon^2 a_1$, like Koiter's, a necessary condition for the robustness of the approximation is that the intersections of the discrete subspaces for the kernel of a_0 are a sequence of dense subspaces for the whole kernel of a_0 ; see Sanchez-Hubert and Sanchez-Palencia (1997), Ch. XI. In the case of the Koiter model, this means that the whole inextensional space $V_F(S)$ (69) can be approximated by the subspaces of the inextensional elements belonging to FE spaces. For hyperbolic shells, the only inextensional elements belonging to FE spaces are zero; see Sanchez-Hubert and Sanchez-Palencia (1997) and Chapelle and Bathe (2003), Section 7.3, which prevents all approximation property of $V_F(S)$ if it is not reduced to $\{0\}$.

This fact is an extreme and general manifestation of the *membrane locking* of shells, also addressed in Pitkäranta (1992) and Gerdes, Matache and Schwab (1998) for cylindrical shells, which are a prototype of shells having a nonzero inextensional space. Plates do not present membrane locking since all elements $\mathbf{z} = (0, \mathbf{z}_3)$ are inextensional, thus can be approximated easily by finite element

subspaces. Nevertheless, as soon as the RM model is used, as can be seen from the structure of the energy (47), a shear locking may appear.

6.2.2 Shear locking of the RM and hierarchical plate models

Shear locking occurs because the FE approximation using C^0 polynomials for the RM family of plates at the limit when $\varepsilon \rightarrow 0$ has to converge to the KL model in energy norm Suri, 2001, requiring C^1 continuity. Let us consider the three-field RM model on the subspace of $V^{\text{RM}}(\Omega^4)$, cf. Section 3.3, of displacements with bending parity: $\{u \in V(\Omega^4), u = (-x_3 \theta_T, z_3)\}$. According to Suri, Babuška and Schwab (1995) we have the following:

Theorem 10. *The p -version of the FEM for the RM plate model without boundary layers, on a mesh of triangles and parallelograms, with polynomial degrees of $p_T \geq 1$ for rotations θ_T and $p_3 \geq p_T$ for z_3 is free of locking in the energy norm.*

For the h -version over a uniform mesh consisting either of triangles or rectangles, to avoid locking the tangential degree p_T has to be taken equal to four or larger, with the transverse degree p_3 being chosen equal to $p_T + 1$. A similar phenomenon was earlier found in connection with 'Poisson Ratio' locking for the equations of elasticity (i.e. conforming elements of degree four or higher encounter no locking); see Scott and Vogelius (1985). In Suri, Babuška and Schwab (1995), it is proven that locking effects (and results) for the $(1,1,2)$ plate model are similar to the RM model because no additional constraints arise as the thickness $\varepsilon \rightarrow 0$. Furthermore, it is stated that locking effects carry over to all hierarchical plate models.

Here we have discussed locking in energy norm. However, if shear stresses are of interest, then locking is significantly worse because these involve an extra power ε^{-1} .

For illustration purposes, consider a *clamped* plate with elliptical midsurface of radii 10 and 5, Young modulus (we recall that the Young modulus is given by $E = \mu(3\lambda + 2\mu)/2(\lambda + \mu)$) and the Poisson ratio by $\nu = \lambda/(2(\lambda + \mu))$, $E = 1$ and Poisson ratio $\nu = 0.3$; see Figure 4. The plate is loaded by a constant pressure of value $(2\varepsilon)^2$.

The discretization is done over a 32 p -element mesh (see Figure 4(a) and (b) for $2\varepsilon = 1$ and 0.1) using two layers, each of dimension ε in the vicinity of the boundary. The FE space is defined with $p_3 = p_T$ ranging from 1 to 8. We show in Figure 5 the locking effects for the RM model with K_{Energy} .

The error plotted in ordinates is the estimated relative discretization error in energy norm between the numerical and exact solution of the RM plate model for each fixed

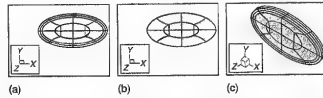


Figure 4. p -FE mesh for $2\varepsilon = 1, 0.1$ for RM model and $2\varepsilon = 1$ for 3-D model. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecom>

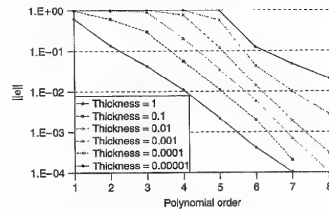


Figure 5. Discretization error versus polynomial degree p for RM plates of various thicknesses ε . A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecom>

thickness ε (it is not the error between the RM numerical solution and the exact 3-D plate model). A similar behavior can be observed with the model $q = (1, 1, 2)$.

To illustrate both the locking effects for the hierarchical family of plates and the modeling errors between the plate models and their 3-D counterpart, we have computed for two thicknesses of plates ($2\varepsilon = 1$ or $2\varepsilon = 0.01$), the solution for the first four plate models (see Table 1 [6]), and for the fully 3-D plate with the degrees $p_T = p_3 = 1, 2, \dots, 8$ with the model represented in Figure 4(c) for $2\varepsilon = 1$.

The relative errors between energy norms of the hierarchical models and the 3D plate model versus the polynomial degree p is shown in Figure 6. As predicted, increasing the order of the plate model does not improve the locking ratio, and as the hierarchical model number is increased the relative error decreases. We note that when

Table 1. Hierarchical plate-model definitions for bending symmetry.

| Model # | 1 (RM) | 2 | 3 | 4 |
|--|---------|---------|---------|---------|
| Degrees $q = (q_1, q_2, q_3)$ | (1,1,0) | (1,1,2) | (3,3,2) | (3,3,4) |
| # independent fields $d = (d_1, d_2, d_3)$ | (1,1,1) | (1,1,2) | (2,2,2) | (2,2,3) |

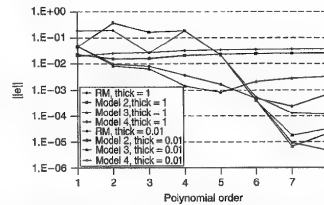


Figure 6. Relative error versus polynomial degree for $2\varepsilon = 1$ and 0.01 for the first 4 hierarchical models. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecom>

$2\varepsilon = 1$ the relative error of the four models converges to the modeling error, which is still quite big since ε is not small, whereas when $2\varepsilon = 0.01$, the error stays larger than 15% for all models when $p \leq 4$, and starts converging for $p \geq 5$.

6.3 Optimal mesh layout for hierarchical models with boundary layers

All hierarchical plate models (besides KL model) exhibit boundary layers. These are rapidly varying components, which decay exponentially with respect to the stretched distance $R = r/\varepsilon$ from the edge, so that at a distance $O(2\varepsilon)$ these are negligible. Finite element solutions should be able to capture these rapid changes. Using the p -version of the finite element method, one may realize exponential convergence rates if a proper design of meshes and selection of polynomial degrees is applied in the presence of boundary layers.

In a 1-D problem with boundary layers, it has been proven in Schwab and Suri (1996) that the p -version over a refined mesh can achieve exponential convergence for the boundary layers, uniformly in ε . The mesh has to be designed so to consist of one $O(p(2\varepsilon))$ boundary layer element at each boundary point. More precisely, the optimal size of the element is $\alpha p(2\varepsilon)$, where, $0 < \alpha < 4/e$ (see Fig. 7).

This result carries over to the heat transfer problem on 2-D domains as shown in Schwab, Suri and Xenophontos (1998), and to the RM plate model, as demonstrated by numerical examples. Typical boundary layer meshes are shown in Figure 4 for $2\varepsilon = 1$ and 0.1. In practice, for ease of computations, two elements in the boundary layer zone are being used, each having the size in the normal direction of ε , independent of the polynomial degree used. This,

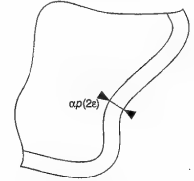


Figure 7. A typical design of the mesh near the boundary for the p -version of the FEM.

although not optimal, still captures well the rapid changes in the boundary layer.

In order to realize the influence of the mesh design over the capture of boundary layer effects, we have again solved numerically the RM plate model for a thickness of $2\varepsilon = 0.01$ and $K_{\text{Deflection}}$ as shear correction factor). Three different mesh layouts have been considered, with two layers of elements in the vicinity of the edge of dimension 0.5, 0.05, and 0.005 (the first two ones are represented in Figure 4). For comparison purposes, we have computed the 3-D solution over a domain having two layers in the thickness direction and two elements in the boundary layer zone of dimension 0.005. We have extracted the vertical displacement u_3 and the shear strain e_{23} along the line starting at $(x_1, x_2) = (9.95, 0)$ and ending at the boundary $(x_1, x_2) = (10, 0)$, that is, in the boundary layer region. Computations use the degrees $p_T = p_3 = 8$. It turns out that the vertical displacement u_3 is rather insensitive to the mesh, whereas the shear strain e_{23} is inadequately computed if the mesh is not properly designed: With the mesh containing fine layers of thickness 0.005, the average relative error is 10%, but this error reaches 100% with mesh layer thickness 0.05 and 400% for the mesh layer thickness 0.5.

Concerning shells, we have seen in Section 4.2 that the Koiter model for clamped elliptic shells admits boundary layers of length scale $\sqrt{\varepsilon}$, and in Section 4.4 that other length scales may appear for different geometries ($\varepsilon^{1/3}$ and $\varepsilon^{1/4}$). Moreover, for Naghdi model, the short length scale ε is also present; see Pitkäranta, Matache and Schwab (2001). Nevertheless, the 'long' length scales $\varepsilon^{1/3}$ and $\varepsilon^{1/4}$ appear to be less frequent. We may expect a similar situation for other hierarchical models. As a conclusion the mesh design for shell of small thicknesses should (at least) take into account both length scales ε and $\sqrt{\varepsilon}$. Another phenomenon should also be considered: Hyperbolic and parabolic shells submitted to a concentrated load or a singular data are expected to propagate singularities along

their zero curvature lines, with the scale width $\varepsilon^{1/3}$; see Pitkäranta, Matache and Schwab (2001).

6.4 Eigen-frequency computations

Eigen-frequency computations are, in our opinion, a very good indicator of (i) the quality of computations, (ii) the nature of the shell (or plate) response. In particular, the bottom of the spectrum indicates the maximal possible stress-strain energy to be expected under a load of given potential energy. From Theorem 7, we may expect that, except in the case of clamped elliptic shells, the ratio between the energy of the response and the energy of the excitation will behave almost as $\mathcal{O}(\varepsilon^{-2})$.

6.4.1 Eigen-frequency of RM versus 3-D for plates

Eigen-frequencies obtained by the p -version finite element method for clamped RM plates and their counterpart 3-D eigen-frequencies have been compared in Dauge and Yosibash (2002), where rectangular plates of dimensions $1 \times 2 \times 2\varepsilon$ have been considered. For isotropic materials

with Poisson coefficient $\nu = 0.3$, the relative error for the first three eigen-frequencies was found negligible (less than 0.12%), for thin plates with slender ratio of less than 1%, and still small (0.2%) for moderately thick plates (slender ratio about 5%).

For some orthotropic materials, much larger relative errors between the RM eigen-frequencies and their 3-D counterparts have been observed even for relatively thin plates. In one of the orthotropic rectangular plate examples in Dauge and Yosibash (2002), for which the boundary layer effect on the eigen-frequencies should be the most pronounced, a very large relative error of 25% has been reported for the first eigen-frequency at $\varepsilon = 0.1$. This is a significant deviation, whereas the RM model underestimates the 'true' 3-D by 25%, and is attributed to the boundary layer effect.

6.4.2 3-D eigen-frequency computations for shells

We present computations on three families of shells, see Figure 8: (a) clamped spherical shells, (b) sensitive spherical shells, (c) flexural cylindrical shells, all with material

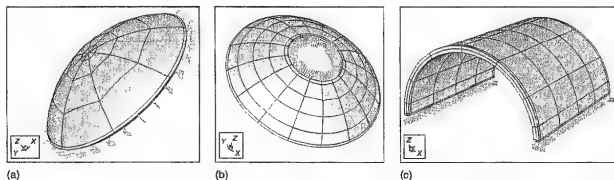


Figure 8. Shell models (a), (b) and (c) for $\varepsilon = 0.04$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

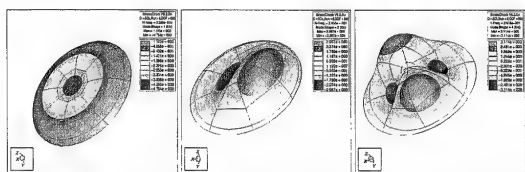


Figure 9. Model (a), vertical components of eigen-modes 1, 2 and 4 for $\varepsilon = 0.08$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

parameters $\nu = 0.3$ and $E = 1$. These three families illustrate the three cases (i), (ii) and (iii) in Theorem 7: The shells (a) are elliptic clamped on their whole boundary, (b) are elliptic, but clamped only on a part of their boundaries, and (c) are parabolic. Note that Theorem 7 states results relating to Koiter eigenvalues and not for 3-D eigenvalues. Nevertheless, a similar behavior can be expected for 3-D eigenvalues.

Family (a). The midsurface S is the portion of the unit sphere described in spherical coordinates by $\varphi \in [0, 2\pi]$ and $\theta \in (\pi/4, \pi/2]$. Thus S is a spherical cap containing the north pole. The family of shells Ω^ε has its upper and lower surfaces characterized by the same angular conditions, and the radii $\rho = 1 + \varepsilon$ and $\rho = 1 - \varepsilon$, respectively. We clamp Ω^ε along its lateral boundary $\theta = \pi/4$.

We have computed the first five eigen-frequencies of the 3-D operator (4) by a FE p -discretization based on 2 layers of elements in the transverse direction and 8×5 elements in the midsurface, including one thin layer of elements in the boundary layer. The vertical (i.e. normal to the tangent plane at the north pole, not transverse to the midsurface) component u_3 for three modes are represented in Figure 9 for the (half)-thickness $\varepsilon = 0.08$. Mode 3 is rotated from mode 2, and mode 5 from mode 4 (double eigen-frequencies). The shapes of the eigen-modes for smaller values of the thickness are similar. Figure 10 provides the three first distinct eigen-frequencies as a function of the thickness in natural scales. In accordance with

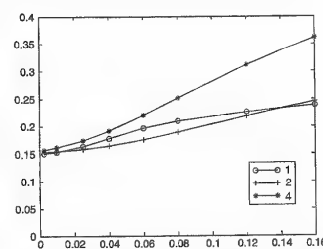


Figure 10. Model (a), Eigen-frequencies versus thickness (2ε). A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

Theorem 7 (i), the smallest eigen-frequencies all tend to the same nonzero limit, which should be the (square root of the) bottom of the membrane spectrum.

Family (b). The midsurface S is the portion of the unit sphere described in spherical coordinates by $\varphi \in [0, 2\pi]$ and $\theta \in (\pi/4, 5\pi/12]$. The family of shells Ω^ε has its upper and lower surfaces characterized by the same angular conditions, and the radii $\rho = 1 + \varepsilon$ and $\rho = 1 - \varepsilon$, respectively.

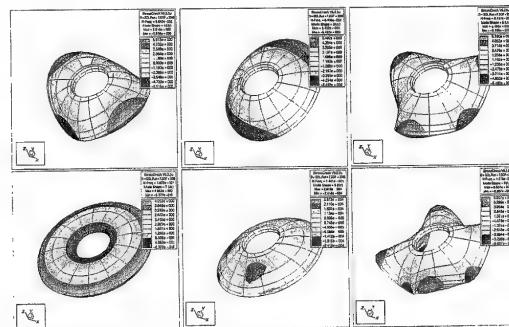


Figure 11. Model (b), Vertical components of modes 1, 3, 5, 7, 8, 9 for $\varepsilon = 0.04$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

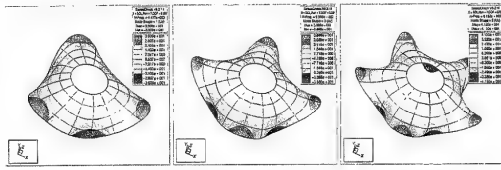


Figure 12. Model (b). Vertical components of modes 1, 3, 5 for $\varepsilon = 0.00125$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

We clamp Ω^* along its lateral boundary $\theta = 5\pi/12$ and let it free along the other lateral boundary $\theta = \pi/4$. This shell is a sensitive one in the sense of Pitkäranta and Sanchez-Palencia (1997), which means that it is sensitive to the thickness and answers differently according to the value of ε .

We have computed the first five (or first ten) eigen-frequencies of the 3-D operator (4) by a FE p -discretization similar to that of (a) (two layers in the transverse direction and 8×4 elements in the surface direction – for the ‘small’ thickness, a globally refined mesh of 16×6 elements has been used). In Figure 11, we plot the vertical components of modes number 1, 3, 5, 7, 8, and 9 for $\varepsilon = 0.04$ and in Figure 12, modes number 1, 3, 5 for $\varepsilon = 0.00125$. In both cases, modes 2, 4, and 6 are similar to modes 1, 3, and 5 respectively and associated with the same (double) eigen-frequencies.

For $\varepsilon = 0.04$, we notice the axisymmetric mode at position 7 (it is at position 5 when $\varepsilon = 0.08$, and 9 for $\varepsilon = 0.02$). Mode 8 looks odd. Indeed, it is very small (less than 10^{-4}) for normalized eigenvectors in $\mathcal{O}(1)$. This means that this mode is mainly supported in its tangential components (we have checked they have a reasonable size). Mode 8 is in fact a *torsion mode*, which means a dominant stretching effect, whereas the other ones have a more pronounced bending character.

Figure 13 provides the first distinct eigen-frequencies classified by the nature of the eigenvector (namely the number of nodal regions of u_3) as a function of the thickness in natural scales. The organization of these eigen-frequencies along affine lines converging to positive limits as $\varepsilon \rightarrow 0$ is remarkable. We may expect a convergence as $\varepsilon \rightarrow 0$ of the solution u^ε of problem (3) provided the loading has a finite number of angular frequencies in φ (the displacement will converge to the highest angular frequency of the load). Nevertheless, such a phenomenon is specific to the axisymmetric nature of the shell (b) and could not be generalized to other sensitive shells. Computations with a

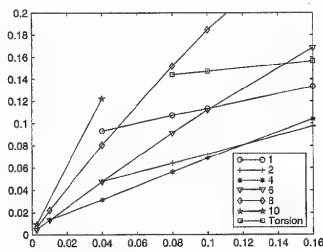


Figure 13. Model (b). Eigen-frequencies versus thickness (2ε). A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

concentrated load (which, of course, has an infinite number of angular frequencies) display a clearly nonconverging behavior (Chapelle and Bathe (2003), Section 4.5.3).

Family (c). The midsurface S is a half-cylinder described in cylindrical coordinates (r, θ, y) by $\theta \in (0, \pi)$, $r = 1$ and $y \in (-1, 1)$. The family of shells Ω^ε has its upper and lower surfaces characterized by the same angular and axial condition, and the radii $r = 1 + \varepsilon$ and $r = 1 - \varepsilon$, respectively. We clamp Ω^ε along its lateral boundaries $\theta = 0$ and $\theta = \pi$ and leave it free everywhere else. This is a well-known example of flexural shell, where the space of inextensional displacements contains the space, cf. (80) (note that, below, $z_r = z_3$)

$$V_{F,0} := \{z = (z_r, z_\theta, z_y); z_y = 0, z_r = z_r(\theta), z_\theta = z_\theta(\theta) \text{ with } \partial_\theta z_\theta = z_r \text{ and } z_\theta = z_r = \partial_\theta z_r = 0 \text{ in } \theta = 0, \pi\} \quad (93)$$

Besides these patterns independent of the axial variable y , there is another subspace $V_{F,1}$ of inextensional displacements, where z_y is independent on y and z_r, z_θ are linear in y :

$$V_{F,1} := \{z = (z_r, z_\theta, z_y); z_y = z_y(\theta), z_\theta = -y \partial_\theta z_y(\theta), z_r = -y \partial_\theta^2 z_y(\theta) \text{ with } z_y = z_\theta = z_r = \partial_\theta z_r = 0 \text{ in } \theta = 0, \pi\} \quad (94)$$

and $V_F = V_{F,0} \oplus V_{F,1}$. We agree to call ‘constant’ the displacements associated with $V_{F,0}$ and ‘linear’ those associated with $V_{F,1}$.

We have computed the first ten eigen-frequencies (4) by a FE p -discretization based on two layers of elements in

the transverse direction and a midsurface mesh of 8×6 curved quadrangles. For the half-thickness $\varepsilon = 0.0025$, we plot the vertical component $u_z = u_3 \sin \theta + u_5 \cos \theta$ of the eigenmodes u : In Figure 14, the first six constant flexural eigenmodes and in Figure 15, the first three linear flexural eigenmodes (their components u_y clearly display a nonzero constant behavior in y). The shapes of the eigen-modes for larger values of the thickness are similar. In Figure 16, we have plotted in logarithmic scale these eigen-frequencies, classified according to the behavior of the flexural eigenmodes (‘constant’ and ‘linear’). The black line has the equation $\varepsilon \mapsto \varepsilon/4$: Thus we can see that the slopes of the eigen-frequency lines are close to 1, as expected by the theory (at least for Koiter model). In Figure 17, we represent the first nonflexural modes (with rank 10 for $\varepsilon = 0.01$ and rank 8, 9 for $\varepsilon = 0.04$).

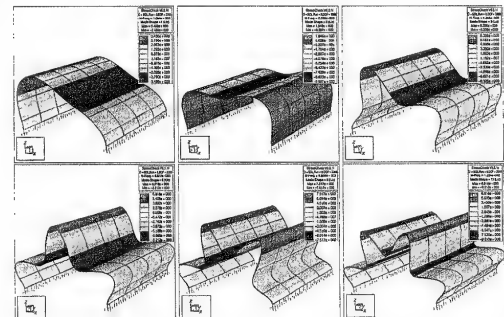


Figure 14. Model (c). Vertical components of modes 1, 2, 5, 6, 9 and 10 for $\varepsilon = 0.0025$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

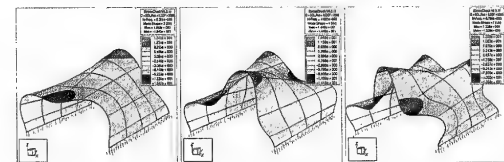


Figure 15. Model (c). Vertical components modes 3, 4 and 7 for $\varepsilon = 0.0025$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

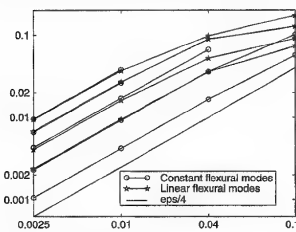


Figure 16. Model (c). Eigen-frequencies versus ϵ in log-log scale. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

6.4.3 Thin element eigen-frequency computations

We present in Tables 2–4 the computation of the first eigen-frequency of the shell Ω^s in families (a) and (b) for a moderate thickness ($\epsilon = 0.02$) and a small thickness ($\epsilon = 0.00125$) and for family (c) for a moderate thickness ($\epsilon = 0.04$) and a small thickness ($\epsilon = 0.0025$), respectively, for a moderate thickness ($\epsilon = 0.04$) and a small thickness

($\epsilon = 0.0025$). The degree q is the degree in the transverse direction (according to Section 6.1.3 there is one layer of elements). We notice that, for an accuracy of 0.01% and $\epsilon = 0.02$, the quadratic kinematics is not sufficient, whereas it is for $\epsilon = 0.00125$. No locking is visible there. In fact, the convergence of the q -models to their own limits is more rapid for $\epsilon = 0.02$.

6.5 Conclusion

It is worthwhile to point out that the most serious difficulties we have encountered in computing all these models occurred for $\epsilon = 0.00125$ and model (b) – the sensitive shell: Indeed, in that case, when $\epsilon \rightarrow 0$, the first eigen-mode is more and more oscillating, and the difficulties of approximation are those of a high-frequency analysis. It is also visible from Tables 3 and 4 that the computational effort is lower for the cylinder than for the sensitive shell, for an even better quality of approximation.

It seems that, considering the high performance of the p -version approximation in a smooth midsurface (for each fixed ϵ and fixed degree q we have an exponential convergence in p), the locking effects can be equilibrated by slightly increasing the degree p as ϵ decreases.

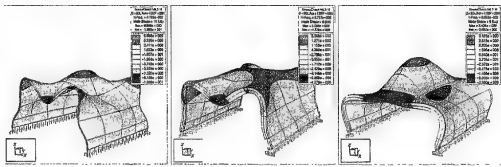


Figure 17. Model (c). First nonflexural modes for $\epsilon = 0.01$ and $\epsilon = 0.04$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

Table 2. Thin element computations for the first eigen-frequency of model (a).

| p | $\epsilon = 0.02$ and $q = 2$ | | | $\epsilon = 0.02$ and $q = 3$ | | | $\epsilon = 0.00125$ and $q = 2$ | | |
|-----|-------------------------------|-----------|--------|-------------------------------|-----------|--------|----------------------------------|-----------|--------|
| | DOF | e-freq. | % err. | DOF | e-freq. | % err. | DOF | e-freq. | % err. |
| 1 | 297 | 0.2271659 | 37.967 | 396 | 0.2264908 | 37.557 | 297 | 0.2055351 | 36.437 |
| 2 | 729 | 0.1694894 | 2.938 | 828 | 0.1694269 | 2.900 | 729 | 0.1560694 | 3.601 |
| 3 | 1209 | 0.1652870 | 0.386 | 1308 | 0.1652544 | 0.366 | 1209 | 0.1537315 | 2.049 |
| 4 | 2145 | 0.1648290 | 0.108 | 2244 | 0.1648001 | 0.090 | 2145 | 0.1517604 | 0.741 |
| 5 | 3321 | 0.1646992 | 0.029 | 3636 | 0.1646693 | 0.011 | 3321 | 0.1508741 | 0.152 |
| 6 | 4737 | 0.1646859 | 0.021 | 5268 | 0.1646555 | 0.002 | 4737 | 0.1506988 | 0.036 |
| 7 | 6393 | 0.1646849 | 0.020 | 7140 | 0.1646544 | 0.002 | 6393 | 0.1506544 | 0.007 |
| 8 | 8289 | 0.1646849 | 0.020 | 9252 | 0.1646543 | 0.002 | 8289 | 0.1506447 | 0.000 |

Table 3. Thin element computations for the first eigen-frequency of model (b).

| p | $\epsilon = 0.02$ and $q = 2$ | | | $\epsilon = 0.02$ and $q = 3$ | | | $\epsilon = 0.00125$ and $q = 2$ | | |
|-----|-------------------------------|-----------|--------|-------------------------------|-----------|--------|----------------------------------|-----------|--------|
| | DOF | e-freq. | % err. | DOF | e-freq. | % err. | DOF | e-freq. | % err. |
| 1 | 864 | 0.0597700 | 89.68 | 1152 | 0.0595287 | 88.91 | 864 | 0.0462144 | 932.2 |
| 2 | 2016 | 0.0326855 | 3.73 | 2304 | 0.0326036 | 3.46 | 2016 | 0.0129819 | 189.9 |
| 3 | 3168 | 0.0318094 | 0.95 | 3456 | 0.0317325 | 0.70 | 3168 | 0.0064504 | 44.06 |
| 4 | 5472 | 0.0316330 | 0.39 | 5760 | 0.0315684 | 0.18 | 5472 | 0.0047030 | 5.04 |
| 5 | 8352 | 0.0316071 | 0.30 | 9216 | 0.0315319 | 0.06 | 8352 | 0.0045085 | 0.69 |
| 6 | 11808 | 0.0316011 | 0.28 | 13248 | 0.0315223 | 0.03 | 11808 | 0.0044800 | 0.06 |
| 7 | 15840 | 0.0316000 | 0.28 | 17856 | 0.0315200 | 0.03 | 15840 | 0.0044780 | 0.01 |
| 8 | 20448 | 0.0315998 | 0.28 | 23040 | 0.0315195 | 0.03 | 20448 | 0.0044779 | 0.01 |

Table 4. Thin element computations for the first eigen-frequency of model (c).

| p | $\epsilon = 0.04$ and $q = 2$ | | | $\epsilon = 0.04$ and $q = 3$ | | | $\epsilon = 0.0025$ and $q = 2$ | | |
|-----|-------------------------------|-----------|--------|-------------------------------|-----------|--------|---------------------------------|-----------|--------|
| | DOF | e-freq. | % err. | DOF | e-freq. | % err. | DOF | e-freq. | % err. |
| 1 | 567 | 0.0514951 | 210.2 | 756 | 0.0510683 | 208.7 | 567 | 0.0397025 | 3666. |
| 2 | 1311 | 0.0207290 | 24.9 | 1500 | 0.0206911 | 24.7 | 1311 | 0.0079356 | 653.1 |
| 3 | 2055 | 0.0167879 | 1.2 | 2244 | 0.0167596 | 0.98 | 2055 | 0.0011505 | 9.188 |
| 4 | 3531 | 0.0166354 | 0.02 | 3720 | 0.0166091 | 0.08 | 3531 | 0.0010578 | 0.395 |
| 5 | 5367 | 0.0166293 | 0.02 | 5928 | 0.0166011 | 0.03 | 5367 | 0.0010548 | 0.108 |
| 6 | 7563 | 0.0166289 | 0.02 | 8496 | 0.0166004 | 0.02 | 7563 | 0.0010541 | 0.045 |
| 7 | 10119 | 0.0166288 | 0.02 | 11424 | 0.0166003 | 0.02 | 10119 | 0.0010538 | 0.012 |
| 8 | 13035 | 0.0166288 | 0.02 | 14712 | 0.0166002 | 0.02 | 13035 | 0.0010537 | 0.002 |

Of course, there exist many strategies to overcome locking in different situations: Let us quote here (Bathe and Brezzi, 1985; Brezzi, Bathe and Fortin, 1989; Arnold and Brezzi, 1997) as 'early references', on mixed methods, which result in a relaxation of the zero-membrane-energy constraint. These methods are addressed in other chapters of the Encyclopedia.

ACKNOWLEDGMENTS

The authors wish to thank Dominique Chapelle (INRIA) for stimulating discussions, Martin Costabel and Yvon Lafranche (zig4tex macro package for drawing figures, see <http://perso.univ-rennes1.fr/yvon.lafranche/zig4tex>) (University of Rennes) for their valuable technical support.

NOTES

- [1] We have a similar situation with plates, where the solution $u^{s,RL}$ of the Kirchhoff–Love model gives back the first generating terms on the asymptotics of u^s , cf. Theorem 2.

- [2] The actual Kirchhoff–Love displacement (satisfying $e_{33} = 0$) is slightly different, containing an extra quadratic surface term.
- [3] The complementing operator \mathbf{C} defined in (45) for plates satisfies $\mathbf{C}u_{kl}^{1,0} = u_{kl}^{1,1,2}$.
- [4] These norms are those of the domains of the fractional powers A^s of the Sturm–Liouville operator $A: \xi \mapsto \partial_x((1-x^2)\partial_x \xi)$ on the interval $(-1, 1)$. Such an approach is now a standard tool in the p -version analysis.
- [5] Of course, different mesh designs are possible on thin domains. If one wants to capture boundary layer terms with an exponential rate of convergence, a h - p refinement should be implemented near the edges of Ω^s , Dauge and Schwab (2002).
- [6] Here, for ease of presentation, we use the numbering system for plate models displayed in Table 1, where we also provide the number d_i of fields in each direction for bending models, that is, for which the surface components are odd and the normal component even in X_3 .

REFERENCES

- Acis RL, Szabo BA and Schwab C. Hierarchic models for laminated plates and shells. *Comput. Methods Appl. Mech. Eng.* 1999; 174(1–4):79–107.

- Agmon S, Douglis A and Nirenberg L. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions II. *Commun. Pure Appl. Math.* 1964; 17:35–92.
- Agratov II and Nazarov SA. Asymptotic analysis of problems in junctions of domains of different limit dimension. An elastic body pierced by thin rods. *J. Math. Sci. (New York)* 2000; 102(5):4349–4387.
- Akian JL and Sanchez-Palencia E. Approximation de coques élastiques minces par facettes planes: Phénomène de blocage membranaire. *C. R. Acad. Sci. Paris, Sér. I* 1992; 315:363–369.
- Andreoli G and Faou E. Complete asymptotics for shallow shells. *Asymptot. Anal.* 2001; 25(3–4):239–270.
- Antic S and Léger A. Formulation bidimensionnelle exacte du modèle de coque 3D de Kirchhoff-Love. *C. R. Acad. Sci. Paris, Sér. I Math.* 1999; 329(8):741–746.
- Arnold DN and Brezzi F. Locking-free finite element methods for shells. *Math. Comput.* 1997; 66(217):1–14.
- Arnold DN and Falk RS. The boundary layer for the Reissner-Mindlin plate model. *SIAM J. Math. Anal.* 1990b; 21(2):281–312.
- Arnold DN and Falk RS. Asymptotic analysis of the boundary layer for the Reissner-Mindlin plate model. *SIAM J. Math. Anal.* 1996; 27(2):486–514.
- Avalishvili M and Gordeziani D. Investigation of two-dimensional models of elastic prismatic shell. *Georgian Math. J.* 2003; 10(1):17–36.
- Babuška I and Li L. Hierarchic modeling of plates. *Comput. Struct.* 1991; 40:419–430.
- Babuška I and Li L. The h-p-version of the finite element method in the plate modelling problem. *Commun. Appl. Numer. Methods* 1992a; 8:17–26.
- Babuška I and Li L. The problem of plate modelling: Theoretical and computational results. *Comput. Methods Appl. Mech. Eng.* 1992b; 100:249–273.
- Babuška I and Pitkäranta J. The plate paradox for hard and soft simple support. *SIAM J. Math. Anal.* 1990; 21:551–576.
- Babuška I and Suri M. On locking and robustness in the finite element method. *SIAM J. Numer. Anal.* 1992; 29:1261–1293.
- Babuška I, d'Harcourt JM and Schwab C. *Optimal Shear Correction Factors in Hierarchic Plate Modelling*. Technical Note BN-1129. Institute for Physical Science and Technology, University of Maryland: College Park, 1991a.
- Babuška I, Szabó BA and Actis RL. Hierarchic models for laminated composites. *Int. J. Numer. Methods Eng.* 1992; 33(3):503–525.
- Bathe KJ and Brezzi F. On the convergence of a four node plate bending element based on Mindlin-Reissner plate theory and a mixed interpolation. In *The Mathematics of Finite Elements and Applications*, vol. 5, Whiteman JR (ed.). Academic Press: London, 1985; 491–503.
- Bernadou M and Ciarlet PG. Sur l'ellipticité du modèle linéaire de coques de W.T. Koiter. In *Computing Methods in Applied Sciences and Engineering*, Lecture Notes in Economics and Mathematical Systems, vol. 134, Glowinski R and Lions JL (eds). Springer-Verlag: Heidelberg, 1976; 89–136.
- Bischoff M and Ramm E. On the physical significance of higher order kinematic and static variables in a three-dimensional shell formulation. *Int. J. Solids Struct.* 2000; 37:6933–6960.
- Brezzi F, Bathe K-J and Fortin M. Mixed-interpolated elements for Reissner-Mindlin plates. *Int. J. Numer. Methods Eng.* 1989; 28(8):1787–1801.
- Budiansky B and Sanders JL. On the "best" first-order linear shell theory. In *Progress in Applied Mechanics*, Anniversary Volume, Prager W (ed.). Macmillan: New York, 1967; 129–140.
- Chapelle D and Bathe K-J. The mathematical shell model underlying general shell elements. *Int. J. Numer. Methods Eng.* 2000; 48(2):289–313.
- Chapelle D and Bathe KJ. *The Finite Element Analysis of Shells – Fundamentals*. Computational Fluid and Solid Mechanics. Springer: Berlin, 2003.
- Chapelle D, Ferent A and Bathe K-J. 3D-shell elements and their underlying mathematical model. *Math. Models Methods Appl. Sci.* 2004; 14(1):105–142.
- Chapelle D, Ferent A and Le Tallec P. The treatment of "pinching locking" in 3D-shell elements. *M2AN Math. Modell. Numer. Anal.* 2003; 37(1):143–158.
- Ciarlet PG. *Mathematical Elasticity, Vol. I, Three-Dimensional Elasticity*. North Holland: Amsterdam, 1988.
- Ciarlet PG. *Mathematical Elasticity, Vol. II, Theory of Plates*. North Holland: Amsterdam, 1997.
- Ciarlet PG. *Mathematical Elasticity*, vol. III. North Holland: Amsterdam, 2000.
- Ciarlet PG and Destuynder P. A justification of the two-dimensional plate model. *J. Méc.* 1979a; 18:315–344.
- Ciarlet PG and Kessavan S. Two-dimensional approximation of three-dimensional eigenvalue problems in plate theory. *Comput. Methods Appl. Mech. Eng.* 1981; 26:149–172.
- Ciarlet PG and Lods V. Asymptotic analysis of linearly elastic shells. I. Justification of membrane shell equations. *Arch. Ration. Mech. Anal.* 1996a; 136:119–161.
- Ciarlet PG and Lods V. Asymptotic analysis of linearly elastic shells. III. Justification of Koiter's shell equations. *Arch. Ration. Mech. Anal.* 1996b; 136:191–200.
- Ciarlet PG, Lods V and Miara B. Asymptotic analysis of linearly elastic shells. II. Justification of flexural shell equations. *Arch. Ration. Mech. Anal.* 1996; 136:163–190.
- Ciarlet PG and Paumier JC. A justification of the Marguerre-von Kármán equations. *Comput. Mech.* 1986; 1:177–202.
- Dauge M and Faou E. *Koiter Estimate Revisited*. Research report, INRIA, 2004; to appear.
- Dauge M and Gruais I. Asymptotics of arbitrary order for a thin elastic clamped plate. I: Optimal error estimates. *Asymptot. Anal.* 1996; 13:167–197.
- Dauge M and Gruais I. Asymptotics of arbitrary order for a thin elastic clamped plate. II: Analysis of the boundary layer terms. *Asymptot. Anal.* 1998a; 16:99–124.
- Dauge M and Schwab C. *hp-FEM for three-dimensional elastic plates*. *M2AN Math. Model. Numer. Anal.* 2002; 36(4):597–630.
- Dauge M and Yosibash Z. Boundary Layer Realization in Thin Elastic 3-D Domains and 2-D Hierarchic Plate Models. *Int. J. Solids Struct.* 2000; 37:2443–2471.
- Dauge M and Yosibash Z. Eigen-frequencies in thin elastic 3-D domains and Reissner-Mindlin plate models. *Math. Methods Appl. Sci.* 2002; 25(1):21–48.
- Dauge M, Gruais I and Rösle A. The influence of lateral boundary conditions on the asymptotics in thin elastic plates. *SIAM J. Math. Anal.* 1999/2000; 31(2):305–345.
- Dauge M, Djurdjevic I, Faou E and Rösle A. Eigenmodes asymptotic in thin elastic plates. *J. Math. Pures Appl.* 1999; 78:925–954.
- Düster A, Bröker H and Rank E. The p-version of the finite element method for three-dimensional curved thin walled structures. *Int. J. Numer. Methods Eng.* 2001; 52:673–703.
- Faou E. Développements asymptotiques dans les coques elliptiques: équations tridimensionnelles linéarisées. *C. R. Acad. Sci. Paris, Sér. I Math.* 2001a; 333(4):389–394.
- Faou E. Développements asymptotiques dans les coques elliptiques: modèle de Koiter. *C. R. Acad. Sci. Paris, Sér. I Math.* 2001b; 333(2):139–143.
- Faou E. Elasticity on a thin shell: Formal series solution. *Asymptot. Anal.* 2002; 31:317–361.
- Faou E. *Multiscale Expansions for Linear Clamped Elliptic Shells*. Research Report RR-4956, INRIA; *Commun. Partial Diff. Equations*, 2004, to appear.
- Friedrichs KO and Dressler RF. A boundary-layer theory for elastic plates. *Commun. Pure Appl. Math.* 1961; 14:1–33.
- Genevey K. A regularity result for a linear membrane shell problem. *RAIRO Modél. Math. Anal. Numér.* 1996; 30(4):467–488.
- Gerdes K, Matache AM and Schwab C. Analysis of membrane locking in hp FEM for a cylindrical shell. *ZAMM Z. Angew. Math. Mech.* 1998; 78(10):663–686.
- Gol'denveizer AL. Derivation of an approximate theory of bending of a plate by the method of asymptotic integration of the equations of the theory of elasticity. *Prikl. Matem. Mekhan* 1962; 26(4):668–686, English translation *J. Appl. Math. Mech.* 1964; 1000–1025.
- Gregory RD and Wan FY. Decaying states of plane strain in a semi-infinite strip and boundary conditions for plate theory. *J. Elastic.* 1984; 14:27–64.
- Havu V and Pitkäranta J. Analysis of a bilinear finite element for shallow shells. I. Approximation of inextensional deformations. *Math. Comput.* 2002; 71(239):923–943.
- Havu V and Pitkäranta J. Analysis of a bilinear finite element for shallow shells. II. Consistency error. *Math. Comput.* 2003; 72(244):1635–1653.
- Il'in AM. *Matching of Asymptotic Expansions of Solutions of Boundary Value Problems*, vol. 102 of *Translations of Mathematical Monographs*. American Mathematical Society: Providence, 1992.
- Irigoien H and Vialto JM. Error estimation in the Bernoulli-Navier model for elastic rods. *Asymptot. Anal.* 1999; 21(1):71–87.
- John F. Refined interior equations for thin elastic shells. *Commun. Pure Appl. Math.* 1971; 24:583–615.
- Koiter WT. A consistent first approximation in the general theory of thin elastic shells. In *Proceedings of IUTAM Symposium on the Theory on Thin Elastic Shells, August 1959*, Delft, 1960; 12–32.
- Koiter WT. On the foundations of the linear theory of thin elastic shells. I. *Proc. Kon. Ned. Akad. Wetensch., Ser. B* 1970a; 73:169–182.
- Koiter WT. On the foundations of the linear theory of thin elastic shells. II. *Proc. Kon. Ned. Akad. Wetensch., Ser. B* 1970b; 73:183–195.
- Koiter WT and Simmonds JG. *Foundations of shell theory. Theoretical and Applied Mechanics*. Springer: Berlin, 1973; 150–176; *Proceedings of Thirteenth International Congress, Moscow University, Moscow*, 1972.
- Kondrat'ev VA. Boundary-value problems for elliptic equations in domains with conical or angular points. *Trans. Moscow Math. Soc.* 1967; 16:227–313.
- Kozlov V, Maz'ya V and Movchan A. *Asymptotic Analysis of Fields in Multi-Structures*. Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, Oxford Science Publications: New York, 1999.
- Lods V and Mandaric C. A justification of linear Koiter and Naghdi's models for totally clamped shell. *Asymptot. Anal.* 2002; 31(3–4):189–210.
- Love AEH. *A Treatise on the Mathematical Theory of Elasticity* (4th edn). Dover Publications: New York, 1944.
- Mandaric C. Asymptotic analysis of linearly elastic shells: error estimates in the membrane case. *Asymptot. Anal.* 1998a; 17:31–51.
- Maz'ya VG, Nazarov SA and Plamenetskii BA. *Asymptotische Theorie elliptischer Randwertprobleme in singulär gestörten Gebieten II. Mathematische Monographien*, Band 83. Akademie Verlag: Berlin, 1991b.
- Mindlin RD. Influence of rotatory inertia and shear on flexural motions of isotropic elastic plates. *J. Appl. Mech.* 1951; 18:31–38.
- Naghdi PM. Foundations of elastic shell theory. *Progress in Solid Mechanics*, vol. 4, North Holland: Amsterdam, 1963; 1–90.
- Naghdi PM. The theory of shells and plates. In *Handbuch der Physik*, vol. VI a/2, Flügge S and Truesdell C (eds). Springer-Verlag: Berlin, 1972; 425–640.
- Nazarov SA. Two-term asymptotics of solutions of spectral problems with singular perturbation. *Math. USSR Sbornik* 1991c; 69(2):307–340.
- Nazarov SA. Justification of the asymptotic theory of thin rods. Integral and pointwise estimates. *J. Math. Sci* 1999; 97(4):4245–4279.
- Nazarov SA. Asymptotic analysis of an arbitrarily anisotropic plate of variable thickness (a shallow shell). *Math. Sbornik* 2000a; 191(7):129–159.
- Nazarov SA. On the asymptotics of the spectrum of a problem in elasticity theory for a thin plate. *Sibirsk. Mat. Zh* 2000b; 41(4):iii, 895–912.
- Nazarov SA and Zorin IS. Edge effect in the bending of a thin three-dimensional plate. *Prikl. Matem. Mekhan* 1989; 53(4):642–650. English translation *J. Appl. Math. Mech.* 1989; 50–507.
- Novozhilov VV. *Thin Shell Theory*. Walters-Noordhoff Publishing: Groningen, 1959.

- Oleinik OA, Shamaev AS and Yosifian GA. *Mathematical Problems in Elasticity and Homogenization. Studies in Mathematics and Its Applications*. North Holland: Amsterdam, 1992.
- Paumier JC. Existence and convergence of the expansion in the asymptotic theory of elastic thin plates. *Math. Modell. Numer. Anal.* 1990; 25(3):371–391.
- Paumier J-C and Raoult A. Asymptotic consistency of the polynomial approximation in the linearized plate theory. Application to the Reissner-Mindlin model. In *Elasticité, Viscoélasticité et Contrôle Optimal* (Lyon, 1995), vol. 2 of ESAIM Proc., Society of Mathematical Applied Industry, Paris, 1997; 203–213.
- Pitkäranta J. The problem of membrane locking in finite element analysis of cylindrical shells. *Numer. Math.* 1992; 61:523–542.
- Pitkäranta J and Sanchez-Palencia E. On the asymptotic behaviour of sensitive shells with small thickness. *C. R. Acad. Sci. Paris, Sér. II* 1997; 325:127–134.
- Pitkäranta J, Matache A-M and Schwab C. Fourier mode analysis of layers in shallow shell deformations. *Comput. Methods Appl. Mech. Eng.* 2001; 190:2943–2975.
- Rösle A, Bischoff M, Wendland W and Ramm E. On the mathematical foundation of the (1,1,2)-plate model. *Int. J. Solids Struct.* 1999; 36(14):2143–2168.
- Sanchez-Hubert J and Sanchez-Palencia E. *Coques élastiques minces. Propriétés asymptotiques, Recherches en mathématiques appliquées*. Masson: Paris, 1997.
- Schwab C. A-posteriori modeling error estimation for hierarchic plate models. *Numer. Math.* 1996; 74(2):221–239.
- Schwab C and Suri M. The p and hp versions of the finite element method for problems with boundary layers. *Math. Comput.* 1996; 65:1403–1429.
- Schwab C and Wright S. Boundary layer approximation in hierarchical beam and plate models. *J. Elastic.* 1995; 38:1–40.
- Schwab C, Suri M and Xenophontos C. The hp finite element method for problems in mechanics with boundary layers. *Comput. Methods Appl. Mech. Eng.* 1998; 157:311–333.
- Scott LR and Vogelius M. Conforming finite element methods for incompressible and nearly incompressible continua. *Large-Scale Computations in Fluid Mechanics*, Part 2 (La Jolla, 1983), vol. 22 of *Lectures in Applied Mathematics*, American Mathematical Society: Providence, 1985; 221–244.
- Stein E and Ohnibus S. Coupled model- and solution-adaptivity in the finite-element method. *Comput. Methods Appl. Mech. Engng.* 1997; 150(1–4):327–350.
- Stoker H. *Differential Geometry. Pure and Applied Mathematics*, vol. XX. Interscience Publishers, John Wiley & Sons: New York-London-Sydney, 1969.
- Suri M. The p and hp finite element method for problems on thin domains. *J. Comput. Appl. Math.* 2001; 128(1–2):235–260.
- Suri M, Babuška I and Schwab C. Locking effects in the finite element approximation of plate models. *Math. Comput.* 1995; 64:461–482.
- Szabó B and Babuška I. *Finite Element Analysis*. Wiley: New York, 1991.
- Szabó B and Sührmann GJ. Hierarchic plate and shell models based on p-extension. *Int. J. Numer. Methods Eng.* 1988; 26:1855–1881.

- Vekua IN. On a method of computing prismatic shells. *Akad. Nauk Grazin. SSR. Trudy Tbiliss. Mat. Inst. Razmadze* 1955; 21:191–259.
- Vekua IN. Theory of thin shallow shells of variable thickness. *Akad. Nauk Grazin. SSR. Trudy Tbiliss. Mat. Inst. Razmadze* 1965; 30:3–103.
- Vekua IN. *Shell Theory: General Methods of Construction. Monographs, Advanced Texts and Surveys in Pure and Applied Mathematics*, 25. Pitman (Advanced Publishing Program): Boston, 1968.
- Vishik MI and Lyusternik LA. Regular degeneration and boundary layers for linear differential equations with small parameter. *Am. Math. Soc. Transl.* 1962; 2(20):239–364.
- Vogelius M and Babuška I. On a dimensional reduction method. III. A posteriori error estimation and an adaptive approach. *Math. Comput.* 1981a; 37(156):361–384.
- Vogelius M and Babuška I. On a dimensional reduction method. II. Some approximation-theoretic results. *Math. Comput.* 1981b; 37(155):47–68.
- Vogelius M and Babuška I. On a dimensional reduction method. I. The optimal selection of basis functions. *Math. Comput.* 1981c; 37(155):31–46.

FURTHER READING

- Agmon S, Douglis A and Nirenberg L. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions I. *Commun. Pure Appl. Math.* 1959; 17:35–92.
- Aganovic C and Tutek Z. A justification of the one-dimensional model of an elastic beam. *Math. Methods Appl. Sci.* 1986; 8:1–14.
- Agranovich MS and Vishik MI. Elliptic problems with a parameter and parabolic problems of general type. *Russian Math. Surv.* 1964; 19:53–157.
- Alessandrini SM, Arnold DN, Falk RS and Madureira AL. Derivation and justification of plate models by variational methods. In *Proceeding of the Summer Seminar of the Canadian Mathematical Society on "Plates and Shells: From Mathematical Theory to Engineering Practice"*, CRM Proceedings and Lecture Notes, Quebec, 1996.
- Andreoli G. Comparaison entre modèles bidimensionnels de coques faiblement courbées. *C. R. Acad. Sci. Paris, Sér. I* 1999b; 329:339–342.
- Andreoli G. *Analyse des coques faiblement courbées*. Thèse de doctorat, Université Pierre et Marie Curie, Paris, 1999a.
- Andreoli G, Dauge M and Faou E. Développement asymptotiques complets pour des coques faiblement courbées encastrées ou libres. *C. R. Acad. Sci. Paris, Sér. I Math.* 2000; 330(6):523–528.
- Angatov II and Nazarov SA. Asymptotic solution to the problem of an elastic body lying on several small supports. *J. Appl. Math. Mech.* 1994; 58(2):303–311.
- Arnold DN and Falk RS. Edge effects in the Reissner-Mindlin plate model. In *Analytical and Computational Models for Shells*,

- Noor AK, Belytschko T and Sino J (eds). *American Society of Mechanical Engineers*: New York, 1990a; 71–90.
- Guo B and Babuška I. Regularity of the solutions for elliptic problems on nonsmooth domains in \mathbb{R}^3 . I. Countably normed spaces on polyhedral domains. *Proc. R. Soc. Edinburgh, Sect. A* 1997b; 127(1):77–126.
- Guo B and Babuška I. Regularity of the solutions for elliptic problems on nonsmooth domains in \mathbb{R}^3 . II. Regularity in neighbourhoods of edges. *Proc. R. Soc. Edinburgh, Sect. A* 1997a; 127(3):517–545.
- Babuška I, d'Harcourt JM and Schwab C. Optimal shear correction factors in hierarchic plate modelling. *Math. Modell. Sci. Comput.* 1991b; 1:1–30.
- Babuška I and Prager M. Reissnerian algorithms in the theory of elasticity. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astr. Phys.* 1960; 8:411–417.
- Basar Y and Kitzig WB. A consistent shell theory for finite deformations. *Acta Mech.* 1988; 76:73–87.
- Bauer I and Reis IL. Nonlinear buckling of rectangular plates. *SIAM J. Appl. Math.* 1965; 13:603–626.
- Berdichevskii VL. *Variatsionnye printsipy mekhaniki sploshnoireduy (Variational Principles of Continuum Mechanics)*, Nauka, Moscow, 1983.
- Berger MS. On the von Kármán equations and the buckling of a thin elastic plate. I. The clamped plate. *Commun. Pure Appl. Math.* 1967; 20:687–719.
- Berger MS and Fife PC. On the von Kármán equations and the buckling of a thin elastic plate. II. Plate with general edge conditions. *Commun. Pure Appl. Math.* 1968; 21:227–241.
- Bermudez A and Viano JM. Une justification déséquations de la thermolasticité des poutres à section variable par des méthodes asymptotiques. *RAIRO Anal. Numér.* 1984; 18:347–376.
- Bernadou M. Variational formulation and approximation of junctions between shells. In *Proceedings, Fifth International Symposium on Numerical Methods in Engineering*, vol. 18 of *Computational Mechanics Publications*, Gruber R. Périaux J and Shaw RP (eds), Springer-Verlag: Heidelberg, 1989; 407–414.
- Bernadou M. *Méthodes d'Éléments Finis pour les Problèmes de Coques Minces*. Masson: Paris, 1994.
- Bernadou M and Boissière JM. *The Finite Element Method in Thin Shell Theory: Application to Arch Dam Simulations*. Birkhäuser: Boston, 1982.
- Bernadou M, Fayolle S and Léné F. Numerical Analysis of junctions between plates. *Comput. Math. Appl. Mech. Eng.* 1989; 74:307–326.
- Bernadou M and Lallane B. Sur l'approximation des coques minces par des méthodes B-splines éléments finis. In *Tendances Actuelles en Calcul des Structures*, Griet JP and Campel GM (eds), Plurails: Paris, 1985; 939–958.
- Bernadou M and Oden JT. An existence theorem for a class of nonlinear shallow shell theories. *J. Math. Pures Appl.* 1981; 60:285–308.
- Bielacki W and Telega JJ. On existence of solutions for geometrically nonlinear shells and plates. *Z. Angew. Math. Mech.* 1988; 68:155–157.
- Blouza A and Le Dret H. Sur le lemme du mouvement rigide. *C. R. Acad. Sci. Paris, Sér. I* 1994b; 319:1015–1020.
- Blouza A and Le Dret H. Existence et unicité pour le modèle de Koiter pour une coque peu régulière. *C. R. Acad. Sci. Paris, Sér. I* 1994a; 319:1127–1132.
- Bolley P, Camus J and Dauge M. Régularité Gevrey pour le problème de Dirichlet dans des domaines à singularités coniques. *Commun. Partial Diff. Equations* 1985; 10(2):391–432.
- Brezzi F and Fortin M. Numerical approximation of Mindlin-Reissner plates. *Math. Comput.* 1986; 47:151–158.
- Busse S, Ciarlet PG and Miara B. Justification d'un modèle linéaire bi-dimensionnel de coques "faiblement courbées" en coordonnées curvilignes. *RAIRO Modél. Math. Anal. Numer.* 1997; 31(3):409–434.
- Chen C. *Asymptotic Convergence Rates for the Kirchhoff Plate Model*. PhD thesis, Pennsylvania State University, 1995.
- Ciarlet PG. *Elasticité tridimensionnelle. Recherches en mathématiques appliquées*. Masson: Paris, 1986.
- Ciarlet PG. *Plates and Junctions in Elastic Multi-Structures: An Asymptotic Analysis*. RMA, vol. 14. Masson and Springer-Verlag: Paris and Heidelberg, 1990.
- Ciarlet PG and Destuynder P. A justification of a nonlinear model in plate theory. *Comput. Methods Appl. Mech. Eng.* 1979b; 17–18:227–258.
- Ciarlet PG and Lods V. Ellipticité des équations membranaires d'une coque uniformément elliptique. *C. R. Acad. Sci. Paris, Sér. I* 1994a; 318:195–200.
- Ciarlet PG and Lods V. Analyse asymptotique des coques linéairement élastiques. I. Coques "membranaires". *C. R. Acad. Sci. Paris, Sér. I* 1994a; 318:863–868.
- Ciarlet PG, Lods V and Miara B. Analyse asymptotique des coques linéairement élastiques. II. Coques en "en flexion". *C. R. Acad. Sci. Paris, Sér. I* 1994; 319:95–100.
- Ciarlet PG and Lods V. Analyse asymptotique des coques linéairement élastiques. III. Une justification du modèle de W. T. Koiter. *C. R. Acad. Sci. Paris, Sér. I* 1994b; 319:299–304.
- Ciarlet PG and Miara B. Justification d'un modèle bi-dimensionnel de coque "peu profonde" élasticités linéarisée. *C. R. Acad. Sci. Paris, Sér. I* 1990; 311:571–574.
- Ciarlet PG and Miara B. Une démonstration simple de l'ellipticité des modèles de coques de W. T. Koiter et de P. M. Naghdi. *C. R. Acad. Sci. Paris, Sér. I* 1991; 312:411–415.
- Ciarlet PG and Miara B. Justification of the two-dimensional equations of a linearly elastic shallow shell. *Commun. Pure Appl. Math.* 1992a; 45(3):327–360.
- Ciarlet PG and Miara B. On the ellipticity of linear shell models. *Z. Angew. Math. Phys.* 1992c; 43:243–253.
- Ciarlet PG and Miara B. Justification of the two-dimensional equations of a linearly elastic shallow shell. *Commun. Pure Appl. Math.* 1992b; 45:327–360.
- Ciarlet PG and Sanchez-Palencia E. Un théorème d'existence et d'unicité pour les équations de coques membranaires. *C. R. Acad. Sci. Paris, Sér. I* 1993; 317:801–805.
- Costabel M and Dauge M. General Edge Asymptotics of Solutions of Second Order Elliptic Boundary Value Problems I. *Proc. R. Soc. Edinburgh* 1993b; 123A:109–155.

- Costabel M and Dauge M. General Edge Asymptotics of Solutions of Second Order Elliptic Boundary Value Problems II. *Proc. R. Soc. Edinburgh* 1993c; 123A:157–184.
- Costabel M and Dauge M. Edge asymptotics on a skew cylinder: complex variable form. *Partial Differential Equations*, Banach Center Publications, vol. 27. Warszawa: Poland, 1992; 81–90.
- Costabel M and Dauge M. Construction of corner singularities for Agmon-Douglis-Nirenberg elliptic systems. *Math. Nachr.* 1993a; 162:209–237.
- Costabel M and Dauge M. Stable asymptotics for elliptic systems on plane domains with corners. *Commun. Partial Diff. Equations* n° 1994; 9 & 10:1677–1726.
- Costabel M and Dauge M. Computation of corner singularities in linear elasticity. In *Boundary Value Problems and Integral Equations in Nonsmooth Domains*, Lecture Notes in Pure and Applied Mathematics, vol. 167, Costabel M, Dauge M and Nicaise S (eds). Marcel Dekker: New York, 1995; 59–68.
- Coutin R. Théorème d'existence et d'unicité pour un problème de coque élastique dans le cas d'un modèle linéaire de Naghdi. *RAIRO Anal. Numér.* 1978; 12:51–57.
- Damlamian A and Vogelius M. Homogenization limits of the equations of elasticity in thin domains. *SIAM J. Math. Anal.* 1987; 18(2):435–451.
- Dauge M and Gruais I. *Complete Asymptotics and Optimal Error Estimates in the Kirchhoff-Love Problem*. Preprint 95-06, Université de Rennes 1, 1995a.
- Dauge M and Gruais I. Edge layers in thin elastic plates. *Comput. Methods Appl. Mech. Eng.* 1998b; 157:335–347.
- Dauge M and Gruais I. Développement asymptotique d'ordre arbitraire pour une plaque élastique mince encastrée. *C. R. Acad. Sci. Paris, Sér. I* 1995b; 321:375–380.
- Deutray R and Lions J-L. *Analyse Mathématique et Calcul Numérique pour les Sciences et les Techniques*. Tome 1. Masson: Paris, 1984.
- Dauge M. *Elliptic Boundary Value Problems in Corner Domains – Smoothness and Asymptotics of Solutions*, Lecture Notes in Mathematics, vol. 1341. Springer-Verlag: Berlin, 1988.
- Dauge M, Djurdjevic I and Rösle A. Higher order bending and membrane responses of thin linearly elastic plates. *C. R. Acad. Sci. Paris, Sér. I* 1998b; 326:519–524.
- Dauge M, Djurdjevic I and Rösle A. Full Asymptotic Expansions for Thin Elastic Free Plates. *C. R. Acad. Sci. Paris, Sér. I* 1998a; 326:1243–1248.
- de Figueiredo I and Trabucho L. A Galerkin approximation for linear elastic shallow shells. *Comput. Mech.* 1992; 10:107–119.
- Delfour MC and Zolésio JP. Differential equations for linear shells: Comparison between intrinsic and classical models. *Adv. Math. Sci.* 1997; 11:42–144.
- Destuynder P. *Sur une Justification des Modèles de Plaques et de Coques par les Méthodes Asymptotiques*. Thèse d'Etat, Université Pierre et Marie Curie, Paris, 1980.
- Destuynder P. Comparaison entre les modèles tridimensionnels et bidimensionnels de plaques élastiques. *RAIRO Anal. Numér.* 1981; 15:331–369.
- Destuynder P. A classification of thin shell theories. *Acta Appl. Math.* 1985; 4:15–63.
- Destuynder P. *Une théorie asymptotique des plaques minces élastiques linéaires*. Masson: Paris, 1986.
- Destuynder P. *Modélisation des coques minces élastiques. Physique fondamentale et appliquée*. Masson: Paris, 1990.
- Dickman M. *Theory of Thin Elastic Shells*. Pitman: Boston, 1982.
- Carmo MP. *Differential Geometry of Curves and Surfaces*. Prentice Hall, 1976.
- Carmo MP. *Riemannian Geometry. Mathematics: Theory and Applications*. Birkhäuser: Boston, 1992.
- Dunford N and Schwartz JT. *Linear Operators – Part II*. Interscience Publishers: New York, 1963.
- Duvaut G and Lions J-L. *Les Inéquations en Mécanique et en Physique*. Dunod: Paris, 1972.
- Duvaut G and Lions J-L. Problèmes unilatéraux dans la théorie de la flexion forte des plaques. *J. Méc.* 1974a; 13:51–74.
- Duvaut G and Lions J-L. Problèmes unilatéraux dans la théorie de la flexion forte des plaques. II: le cas d'évolution. *J. Méc.* 1974b; 13:245–266.
- Ekhaus W. Boundary layers in linear elliptic singular perturbations. *SIAM Rev.* 1972; 14:225–270.
- Ekhaus W. *Asymptotic Analysis of Singular Perturbations*. North Holland: Amsterdam, 1979.
- Faou E. Élasticité linéarisée tridimensionnelle pour une coque mince: résolution en série formelle en puissances de l'épaisseur. *C. R. Acad. Sci. Paris, Sér. I Math.* 2000b; 330(5):415–420.
- Faou E. *Développements asymptotiques dans les coques linéairement élastiques*. Thèse, Université de Rennes 1, 2000a.
- Feigin VI. Elliptic equations in domains with multidimensional singularities of the boundary. *Uspehi-Mat. Nauk.* 1972; 2:183, 184.
- Fichera G. Existence theorems in elasticity. In *Handbuch der Physik*, vol. VIa-2, Flügge S and Truesdell C (eds). Springer-Verlag: Berlin, 1972; 347–389.
- Friedrichs KO and Stokes JJ. Buckling of a circular plate beyond the critical thrust. *J. Appl. Mech.* 1942; 9:A7–A14.
- Gol'denveizer AL. *Theory of Elastic Thin Shells*. Pergamon: New York, 1961.
- Gol'denveizer AL. The construction of an approximate theory of shells by means of asymptotic integration of elasticity equations. *Prikl. Math. Mech.* 1963; 27(4):593–608, English translation *J. Appl. Math. Mech.* 1963; 903–924.
- Golevzev AL. The principles of reducing three-dimensional problems of elasticity to two-dimensional problems of the theory of plates and shells. In *Proceedings of the 11th International Congress of Theoretical and Applied Mechanics*, Görtler H (ed.). Springer-Verlag: Berlin, 1964; 306–311.
- Gordeziani DG. The solvability of certain boundary value problems for a variant of the theory of thin shells. *Dokl. Akad. Nauk SSSR* 1974b; 215:1289–1292.
- Gordeziani DG. The accuracy of a certain variant of the theory of thin shells. *Dokl. Akad. Nauk SSSR* 1974a; 216:751–754.
- Gruais I. Modélisation de la jonction entre une plaque et une poutre en élasticité linéarisée. *Modell. Math. Anal. Numér.* 1993b; 27:77–109.
- Gruais I. Modeling of the junction between a plate and a rod in nonlinear elasticity. *Asymptot. Anal.* 1993a; 7:179–194.
- John F. Estimates for the derivatives of the stresses in a thin shell and interior shell equations. *Commun. Pure Appl. Math.* 1965; 18:235–267.
- John F. A priori estimates, geometric effects and asymptotic behaviour. *Commun. Pure Appl. Math.* 1975; 81:1013–1023.
- Kato T. *Perturbation Theory for Linear Operators*. Springer-Verlag: Berlin – Heidelberg – New York, 1976.
- Kato T. Perturbation theory for nullity, deficiency and other quantities of linear operators. *J. Anal. Math.* 1958; 6:261–322.
- Keller HB, Keller JB and Reiss E. Buckled states of circular plates. *Q. J. Appl. Math.* 1962; 20:5–65.
- Kirchhoff G. Über das Gleichgewicht und die Bewegung einer elastischen Scheibe. *J. Reine Angew. Math.* 1850; 40:51–58.
- Kirchhoff G. *Vorlesungen über Mathematische Physik. Mechanik*. Leipzig, 1876.
- Koiter WT. On the nonlinear theory of thin shells. *Proc. Kon. Ned. Akad. Wetensch., Ser. B* 1966; 69:1–59.
- Koiter WT. General theory of shell stability. Thin shell theory. *New Trends and Applications*, vol. 240 of C. I. S. M. Courses and Lectures. Springer-Verlag: New York, 1980; 65–87.
- Kondrat'ev VA. On estimates of intermediate derivatives by means of moments of the highest derivative. *Proc. Steklov Inst. Math.* 1992; 2:193–203.
- Kondrat'ev VA and Oleinik OA. Boundary-value problems for partial differential equations in non-smooth domains. *Russian Math. Surv.* 1983; 38:1–86.
- Kondrat'ev VA and Oleinik OA. On Korn's Inequalities. *C. R. Acad. Sci. Paris, Sér. I* 1989; 308:483–487.
- Kozlov VA, Maz'ya VG and Movchan AB. Asymptotic analysis of a mixed boundary value problem in a multi-structure. *Asymptot. Anal.* 1994; 8:105–143.
- Kozlov VA, Maz'ya VG and Schwab C. On singularities of solutions of the displacement problem of linear elasticity near the vertex of a cone. *Arch. Ration. Mech. Anal.* 1992; 119:197–227.
- Leino Y and Pitkäranta J. On the membrane locking of h - p finite elements in a cylindrical shell problem. *Int. J. Numer. Methods Eng.* 1994; 37(6):1053–1070.
- Lions J-L and Magenes E. *Problèmes aux limites non homogènes et applications*. Dunod: Paris, 1968.
- Lions J-L and Sanchez-Palencia E. Problèmes aux limites sensitifs. *C. R. Acad. Sci. Paris, Sér. I* 1994; 319:1021–1026.
- Lods V and Mandare C. Asymptotic justification of the Kirchhoff-Love assumptions for a linearly elastic clamped shell. *J. Elastic.* 2000; 58(2):105–154.
- Lods V and Mandare C. Error estimates between the linearized three-dimensional shell equations and Naghdi's model. *Asymptot. Anal.* 2001; 28(1):1–30.
- Lods V and Mandare C. Justification asymptotique des hypothèses de Kirchhoff-Love pour une coque encastrée linéairement élastique. *C. R. Acad. Sci. Paris, Sér. I* 1998; 326:909–912.
- Mandare C. Estimations d'erreur dans l'analyse asymptotique des coques linéairement élastiques. *C. R. Acad. Sci. Paris, Sér. I* 1996; 322:895–899.
- Mandare C. Two-dimensional models of linearly elastic shells: error estimates between their solutions. *Math. Mech. Solids* 1998b; 3:303–318.
- Maz'ya VG, Nazarov SA and Plamenetskii BA. *Asymptotische Theorie elliptischer Randwertprobleme in singular gestörten Gebieten I. Mathematische Monographien*, Band 82. Akademie Verlag: Berlin, 1991a.
- Miara B. Optimal spectral approximation in linearized plate theory. *Applicable Anal.* 1989; 31:291–307.
- Miara B. Justification of the asymptotic analysis of elastic plate. I: The linear case. *Asymptot. Anal.* 1994a; 9:47–60.
- Miara B. Justification of the asymptotic analysis of elastic plate. II: The nonlinear case. *Asymptot. Anal.* 1994b; 9:119–134.
- Miara B and Trabucho L. A Galerkin spectral approximation in linearized beam theory. *Modell. Math. Anal. Numér.* 1992; 26:425–446.
- Mielke A. Normal hyperbolicity of center manifolds and Saint-Venant's principle. *Arch. Ration. Mech. Anal.* 1990; 110:353–372.
- Mielke A. Reduction of PDEs in domains with several unbounded directions: a step towards modulation equations. *Z. Angew. Math. Phys.* 1992; 43(3):449–470.
- Mielke A. On the justification of plate theories in linear elasticity theory using exponential decay estimates. *J. Elastic.* 1995; 38:165–208.
- Morgenstern D. Herleitung der Plattentheorie aus der dreidimensionalen Elastizitätstheorie. *Arch. Ration. Mech. Anal.* 1959; 4:145–152.
- Narasimhan R. *Analysis on Real and Complex Manifolds. Advanced Studies in Pure Mathematics*. North Holland: Amsterdam, 1968.
- Nazarov SA. Justification of asymptotic expansions of the eigenvalues of nonselfadjoint singularly perturbed elliptic boundary value problems. *Math. USSR Sbornik* 1987; 57(2):317–349.
- Nazarov SA. On three-dimensional effects near the vertex of a crack in a thin plate. *J. Appl. Math. Mech.* 1991a; 55(4):407–415.
- Nazarov SA. The spatial structure of the stress field in the neighbourhood of the corner point of a thin plate. *J. Appl. Math. Mech.* 1991b; 55(4):523–530.
- Nazarov SA and Plamenetskii BA. *Elliptic Problems in Domains with Piecewise Smooth Boundaries. De Gruyter Expositions in Mathematics*. Walter de Gruyter: Berlin, New York, 1994.
- Necas J and Hlavacek I. *Mathematical Theory of Elastic and Elasto-Plastic Bodies: An Introduction*. Elsevier Scientific Publishing Company: Amsterdam, 1981.
- Paumier JC and Rao B. Qualitative and quantitative analysis of buckling of shallow shells. *Eur. J. Mech. A, Solids* 1989; 8:461–489.
- Piła J and Pitkäranta J. Energy estimates relating different linear elastic models of a thin cylindrical shell I. The membrane-dominated case. *SIAM J. Math. Anal.* 1993; 24(1):1–22.
- Piła J and Pitkäranta J. Energy estimates relating different linear elastic models of a thin cylindrical shell II. The case of free boundary. *SIAM J. Math. Anal.* 1995; 26(4):820–849.

- Pitkäranta J, Leino Y, Ovaskainen O and Piila J. Shell deformation states and the finite element method: a benchmark study of cylindrical shells. *Comput. Methods Appl. Mech. Eng.* 1995; **128**(1–2):81–121.
- Prager W and Synge JL. Approximations in elasticity based on the concept of the function space. *Q. Appl. Math.* 1947; **5**:241–269.
- Rachewski PK. *Riemannsche Geometrie und Tensoranalysis*. VEB deutscher Verlag der Wissenschaften: Berlin, 1959.
- Reissner E. On the theory of bending of elastic plates. *J. Math. Phys.* 1944; **23**:184–191.
- Reissner E. The effect of transverse shear deformations on the bending of elastic plates. *J. Appl. Mech.* 1945; **12**:A69–A77.
- Reissner E. On a variational theorem in elasticity. *J. Math. Phys.* 1950; **28**:90–93.
- Reissner E. On the derivation of boundary conditions for plate theory. *Proc. R. Soc. Ser. A* 1963; **276**:178–186.
- Reissner E. Reflections on the theory of elastic plates. *Appl. Mech. Rev.* 1983; **38**:453–464.
- Reissner E. On small finite deflections of shear deformable elastic plates. *Comput. Methods Appl. Mech. Eng.* 1986; **59**:227–233.
- Rodriguez JM and Viano JM. Analyse asymptotique de l'équation de Poisson dans un domaine mince. Application à la théorie de torsion des poutres élastiques à profil mince. I. Domaine "sans jonctions". *C. R. Acad. Sci. Paris, Sér. I* 1993a; **317**:423–428.
- Rösle A. *Asymptotische Entwicklungen für dünne Platten im Rahmen der linearen Elastostatik*. Doctoral Dissertation, Mathematisches Institut A, Universität Stuttgart, Germany, 1999.
- Sanchez-Hubert J and Sanchez-Palencia E. *Vibration and Coupling of Continuous Systems: Asymptotic Methods*. Springer-Verlag: Heidelberg, 1989.
- Sanchez-Palencia E. *Non-Homogeneous Media and Vibration Theory*, vol. 127 of *Lecture Notes in Physics*. Springer-Verlag: Heidelberg, 1980.
- Sanchez-Palencia E. Forces appliquées à une petite région de surface d'un corps élastique. Applications aux jonctions. *C. R. Acad. Sci. Paris, Sér. II* 1988; **307**:689–694.
- Sanchez-Palencia E. Statique et Dynamique des Coques minces. I. Cas de flexion pure non inhibée. *C. R. Acad. Sci. Paris, Sér. I* 1989a; **309**:411–417.
- Sanchez-Palencia E. Statique et Dynamique des Coques minces. II. Cas de flexion pure inhibée. Approximation Membranaire. *C. R. Acad. Sci. Paris, Sér. I* 1989b; **309**:531–537.
- Sanchez-Palencia E. Passage à la limite de l'élasticité tridimensionnelle à la théorie asymptotique des coques minces. *C. R. Acad. Sci. Paris, Sér. II* 1990b; **311**:909–916.
- Sanchez-Palencia E and Suquet P. Friction and Homogenization of a Boundary. In *Free Boundary Problems: Theory and Applications*, Passano A and Primicerio M (eds). Pitman: London, 1983; 561–571.
- Sanders JL. *An Improved First-Approximation Theory for Thin Shells*. Report 24, NASA, 1959.
- Sändig AM, Richter U and Sändig R. The regularity of boundary value problems for the Lamé equations in a polygonal domain. *Rostock. Math. Kolloq* 1989; **36**:21–50.
- Schwab C. *The Dimension Reduction Method*. PhD thesis, University of Maryland, College Park, 1989.
- Schwab C. Boundary layer resolution in hierarchical models of laminated composites. *Math. Modell. Numer. Anal.* 1994; **28**(5):517–537.
- Schwab Ch. Theory and applications in solid and fluid mechanics. *p- and hp-Finite Element Methods*. The Clarendon Press Oxford University Press: New York, 1998.
- (1996) *Stress Check User's Manual*, Release 2.0. ESRD, St. Louis.
- Shoikhet BA. On asymptotically exact equations of thin plates of complex structures. *Prikl. Matem. Mekhan* 1973; **37**(5):914–924. English translation *J. Appl. Math. Mech.* 1973; **867**–877.
- Shoikhet BA. On existence theorems in linear shell theory. *Prikl. Matem. Mekhan* 1974; **38**(3):567–571. English translation *J. Appl. Math. Mech.* 1974; **527**–531.
- Shoikhet BA. An energy identity in physically nonlinear elasticity and error estimates of the plate equations. *Prikl. Matem. Mekhan* 1976; **40**(2):317–326. English translation *J. Appl. Math. Mech.* 1976; **291**–301.
- Slicaru SL. Sur l'ellipticité de la surface moyenne d'une coque. *C. R. Acad. Sci. Paris, Sér. I* 1996; **322**:97–100.
- Slicaru SL. *Quelques Résultats dans la Théorie des Coques Linéairement Élastiques à Surface Moyenne Uniformément Elliptiques ou Compacte sans bord*. Thèse de doctorat, Université Pierre et Marie Curie, Paris, 1998.
- Timoshenko S and Woinowsky-Krieger W. *Theory of Plates and Shells*. McGraw-Hill: New York, 1959.
- Trabucho L and Viao JM. A derivation of generalized Saint-Venant's torsion theory from three-dimensional elasticity by asymptotic expansion methods. *Appl. Anal.* 1988; **31**:129–148.
- Trabucho L and Viao JM. Existence and characterization of higher order terms in an asymptotic expansion method for linearized elastic beams. *Asymptot. Anal.* 1989; **2**:223–255.
- Trabucho L and Viao JM. Mathematical modeling of rods. In *Handbook of Numerical Analysis*, vol. 4, Ciarlet PG and Lions J-L (eds). North Holland: Amsterdam, 1994.
- Triebel H. *Interpolation Theory, Function Spaces, Differential Operators*. North Holland Mathematical Library, North Holland: Amsterdam, 1978.
- Vishik MI and Lyusternik LA. Asymptotic behaviour of solutions of linear differential equations with large or quickly changing coefficients and boundary condition. *Russian Math. Surv.* 1960; **4**:23–92.

Chapter 9

Mixed Finite Element Methods

Ferdinando Auricchio, Franco Brezzi and Carlo Lovadina

Università di Pavia and IMATI-C.N.R., Pavia, Italy

| | |
|---|-----|
| 1 Introduction | 237 |
| 2 Formulations | 238 |
| 3 Stability of Saddle-Points in Finite Dimensions | 246 |
| 4 Applications | 257 |
| 5 Techniques for Proving the Inf-Sup Condition | 269 |
| 6 Related Chapters | 276 |
| References | 276 |

1 INTRODUCTION

Finite element method is a well-known and highly effective technique for the computation of approximate solutions of complex boundary value problems. Started in the fifties with milestone papers in a structural engineering context (see e.g. references in Chapter 1 of Zienkiewicz and Taylor (2000a) as well as classical references such as Turner *et al.* (1956) and Clough (1965)), the method has been extensively developed and studied in the last 50 years (Bathe, 1996; Brezzi and Fortin, 1991; Becker, Carey and Oden, 1981; Brenner and Scott, 1994; Crisfield, 1986; Hughes, 1987; Johnson, 1992; Ottosen and Petersson, 1992; Quarteroni and Valli, 1994; Reddy, 1993; Wait and Mitchell, 1985) and it is currently used also for the solution of complex nonlinear problems (Bathe, 1996; Bonet and Wood, 1997; Belytschko, Liu and Moran, 2000; Crisfield, 1991; Crisfield, 1997; Simo and Hughes, 1998; Simo,

1999; Zienkiewicz and Taylor, 2000b; Zienkiewicz and Taylor, 2000c).

Within such a broad approximation method, we focus on the often-called *mixed finite element methods*, where in our terminology the word 'mixed' indicates the fact that the problem discretization typically results in a linear algebraic system of the general form

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} \quad (1)$$

with \mathbf{A} and \mathbf{B} matrices and with \mathbf{x} , \mathbf{y} , \mathbf{f} and \mathbf{g} vectors. Also, on mixed finite elements, the bibliography is quite large, ranging from classical contributions (Auluri, Gallagher and Zienkiewicz, 1983; Carey and Oden, 1983; Strang and Fix, 1973; Zienkiewicz *et al.*, 1983) to more recent references (Bathe, 1996; Belytschko, Liu and Moran, 2000; Bonet and Wood, 1997; Brezzi and Fortin, 1991; Hughes, 1987; Zienkiewicz and Taylor, 2000a; Zienkiewicz and Taylor, 2000c). An impressive amount of work has been devoted to a number of different stabilization techniques, virtually for all applications in which mixed formulations are involved. Their treatment is, however, beyond the scope of this chapter, and we will just say a few words on the general idea in Section 4.2.5.

In particular, the chapter is organized as follows. Section 2 sketches out the fact that several physical problem formulations share the same algebraic structure (1), once a discretization is introduced. Section 3 presents a simple, algebraic version of the abstract theory that rules most applications of mixed finite element methods. Section 4 gives several examples of efficient mixed finite element methods. Finally, in Section 5 we give some hints on how to perform a stability and error analysis, focusing on a representative problem (i.e. the Stokes equations).

2 FORMULATIONS

The goal of the present section is to point out that a quite large set of physical problem formulations shares the same algebraic structure (1), once a discretization is introduced.

To limit the discussion, we focus on *steady state* field problems defined in a domain $\Omega \subset \mathbb{R}^d$, with d the Euclidean space dimension. Moreover, we start from the simplest class of physical problems, that is, the one associated to diffusion mechanisms. Classical problems falling in this frame and frequently encountered in engineering are heat conduction, distribution of electrical or magnetic potentials, irrotational flow of ideal fluids, torsion or bending of cylindrical beams.

After addressing the thermal diffusion, as representative of the whole class, we move to more complex problems, such as the steady state flow of an incompressible Newtonian fluid and the mechanics of elastic bodies. For each problem, we briefly describe the local differential equations and possible variational formulations.

Before proceeding, we need to comment on the adopted notation. In general, we indicate scalar fields with nonbold lower-case roman or nonbold lower-case greek letters (such as a, α, b, β), vector fields with bold lower-case roman letters (such as \mathbf{a}, \mathbf{b}), second-order tensors with bold lower-case greek letters or bold upper-case roman letters (such as $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{A}, \mathbf{B}$), fourth-order tensors with upper-case blackboard roman letters (such as \mathbf{D}). We however reserve the letters \mathbf{A} and \mathbf{B} for 'composite' matrices (see e.g. equation (31)). Moreover, we indicate with $\mathbf{0}$ the null vector, with \mathbf{I} the identity second-order tensor and with \mathbf{I} the identity fourth-order tensor.

Whenever necessary or useful, we may use the standard indicial notation to represent vectors or tensors. Accordingly, in a Euclidean space with base vectors \mathbf{e}_i , a vector \mathbf{a} , a second-order tensor $\boldsymbol{\alpha}$, and a fourth-order tensor \mathbf{D} have the following components

$$\begin{aligned} a_i &= \mathbf{a} \cdot \mathbf{e}_i, & \alpha_{ij} &= \boldsymbol{\alpha} \cdot (\mathbf{e}_i \otimes \mathbf{e}_j) \\ \mathbf{D}_{ijkl} &= \mathbf{D}(\mathbf{e}_i \otimes \mathbf{e}_j) : [\mathbf{D}(\mathbf{e}_k \otimes \mathbf{e}_l)] \end{aligned} \quad (2)$$

where \cdot, \otimes , and $:$ indicate respectively the scalar vector product, the (second-order) tensorial vector product, and the scalar (second-order) tensor product. Sometimes, the scalar vector product will be also indicated as $\mathbf{a}^T \mathbf{b}$, where the superscript T indicates transposition.

During the discussion, we also introduce standard differential operators such as gradient and divergence, indicated respectively as '∇' and 'div', and acting either on scalar, vector, or tensor fields. In particular, we have

$$\begin{aligned} \nabla a_i &= a_{i,j}, & \nabla a_{ij} &= a_{i,j} \\ \text{div } \mathbf{a} &= a_{i,i}, & \text{div } \alpha_{ij} &= \alpha_{ij,j} \end{aligned} \quad (3)$$

where repeated subscript indices imply summation and where the subscript comma indicates derivation, that is, $a_{i,j} = \partial a / \partial x_j$.

Finally, given for example, a scalar field a , a vector field \mathbf{a} , and a tensor field $\boldsymbol{\alpha}$, we indicate with δa , $\delta \mathbf{a}$, $\delta \boldsymbol{\alpha}$ the corresponding variation fields and with a^h , \mathbf{a}^h , $\boldsymbol{\alpha}^h$ the corresponding interpolations, expressed in general as

$$a^h = N_k^a \hat{a}_k, \quad \mathbf{a}^h = N_k^a \hat{\mathbf{a}}_k, \quad \boldsymbol{\alpha}^h = N_k^a \hat{\boldsymbol{\alpha}}_k \quad (4)$$

where N_k^a , $N_k^{\mathbf{a}}$, and $N_k^{\boldsymbol{\alpha}}$ are a set of interpolation functions (i.e. the so-called *shape functions*), while \hat{a}_k and $\hat{\mathbf{a}}_k$ are a set of interpolation parameters (i.e. the so-called *degrees of freedom*); clearly, N_k^a , $N_k^{\mathbf{a}}$, and $N_k^{\boldsymbol{\alpha}}$ are respectively scalar, vector, and tensor predefined (assigned) fields, while \hat{a}_k and $\hat{\mathbf{a}}_k$ are scalar quantities, representing the effective unknowns of the approximated problems. With the adopted notation, it is now simple to evaluate the differential operators (3) on the interpolated fields, that is,

$$\begin{aligned} \nabla a^h &= (\nabla N_k^a) \hat{a}_k, & \nabla \mathbf{a}^h &= (\nabla N_k^{\mathbf{a}}) \hat{\mathbf{a}}_k \\ \text{div } \mathbf{a}^h &= (\text{div } N_k^{\mathbf{a}}) \hat{\mathbf{a}}_k, & \text{div } \boldsymbol{\alpha}^h &= (\text{div } N_k^{\boldsymbol{\alpha}}) \hat{\boldsymbol{\alpha}}_k \end{aligned} \quad (5)$$

or in indicial notation

$$\begin{aligned} \nabla a^h|_i &= N_{k,i}^a \hat{a}_k, & \nabla \mathbf{a}^h|_{ij} &= N_{k,ij}^{\mathbf{a}} \hat{\mathbf{a}}_k \\ \text{div } \mathbf{a}^h &= N_{k,i}^{\mathbf{a}}|_i \hat{\mathbf{a}}_k, & \text{div } \boldsymbol{\alpha}^h|_i &= N_{k,i}^{\boldsymbol{\alpha}}|_i \hat{\boldsymbol{\alpha}}_k \end{aligned} \quad (6)$$

2.1 Thermal diffusion

The physical problem

Indicating with θ the body temperature, \mathbf{e} the temperature gradient, \mathbf{q} the heat flux, and with b the assigned heat source per unit volume, a steady state thermal problem in a domain Ω can be formulated as a $(\theta, \mathbf{e}, \mathbf{q})$ three field problem as follows:

$$\begin{cases} \text{div } \mathbf{q} + b = 0 & \text{in } \Omega \\ \mathbf{q} = -\mathbf{D}\mathbf{e} & \text{in } \Omega \\ \mathbf{e} = \nabla \theta & \text{in } \Omega \end{cases} \quad (7)$$

which are respectively the balance equation, the constitutive equation, the compatibility equation.

In particular, we assume a linear constitutive equation (known as Fourier law), where \mathbf{D} is the conductivity material-dependent second-order tensor, in the simple case of thermally isotropic material, $\mathbf{D} = k\mathbf{I}$ with k the isotropic thermal conductivity.

Equation (7) is completed by proper boundary conditions. For simplicity, we consider only the case of trivial

essential conditions on the whole domain boundary, that is,

$$\theta = 0 \quad \text{on } \partial\Omega \quad (8)$$

This position is clearly very restrictive from a physical point of view but it is still adopted since it simplifies the forthcoming discussion, at the same time without limiting our numerical considerations.

As classically done, the three field problem (7) can be simplified eliminating the temperature gradient \mathbf{e} , obtaining a (θ, \mathbf{q}) two field problem

$$\begin{cases} \text{div } \mathbf{q} + b = 0 & \text{in } \Omega \\ \mathbf{q} = -\mathbf{D}\nabla \theta & \text{in } \Omega \end{cases} \quad (9)$$

and the two field problem (9) can be further simplified eliminating the thermal flux \mathbf{q} (or eliminating the fields \mathbf{e} and \mathbf{q} directly from equation (7)), obtaining a θ single field problem

$$-\text{div}(\mathbf{D}\nabla \theta) + b = 0 \quad \text{in } \Omega \quad (10)$$

For the case of an isotropic and homogeneous body, this last equation specializes as follows

$$-k\Delta \theta + b = 0 \quad \text{in } \Omega \quad (11)$$

where Δ is the standard Laplace operator.

Variational principles

The single field equation (10) can be easily derived starting from the potential energy functional

$$\Pi(\theta) = \frac{1}{2} \int_{\Omega} [\nabla \theta \cdot \mathbf{D} \nabla \theta] \, d\Omega + \int_{\Omega} \theta b \, d\Omega \quad (12)$$

Requiring the stationarity of potential (12), we obtain

$$\delta \Pi(\theta) = \int_{\Omega} [(\nabla \delta \theta) \cdot \mathbf{D} \nabla \theta] \, d\Omega + \int_{\Omega} \delta \theta b \, d\Omega = 0 \quad (13)$$

where $\delta \theta$ indicates a possible variation of the temperature field θ and $\delta \Pi(\theta)[\delta \theta]$ indicates the potential variation evaluated at θ in the direction $\delta \theta$. Since functional (12) is convex, we may note that the stationarity requirement is equivalent to a minimization.

Recalling equation (4), we may now introduce an interpolation for the temperature field in the form

$$\theta \approx \theta^h = N_k^{\theta} \hat{\theta}_k \quad (14)$$

as well as a similar approximation for the corresponding variation field, such that equation (13) can be rewritten in

matricial form as follows

$$\mathbf{A} \hat{\boldsymbol{\theta}} = \mathbf{f} \quad (15)$$

with

$$\begin{cases} \mathbf{A}|_{ij} = \int_{\Omega} [\nabla N_i^{\theta} \cdot \mathbf{D} \nabla N_j^{\theta}] \, d\Omega, & \hat{\boldsymbol{\theta}}|_j = \hat{\theta}_j \\ \mathbf{f}|_i = - \int_{\Omega} [N_i^{\theta} b] \, d\Omega \end{cases} \quad (16)$$

Besides the integral form (13) associated to the single field equation (10), it is also possible to associate an integral form to the two field equation (9) starting now from the more general *Hellinger-Reissner functional*

$$\begin{aligned} \Pi^{\text{HR}}(\theta, \mathbf{q}) &= -\frac{1}{2} \int_{\Omega} [\mathbf{q} \cdot \mathbf{D}^{-1} \mathbf{q}] \, d\Omega - \int_{\Omega} [\mathbf{q} \cdot \nabla \theta] \, d\Omega \\ &\quad + \int_{\Omega} \theta b \, d\Omega \end{aligned} \quad (17)$$

Requiring the stationarity of functional (17), we obtain

$$\begin{cases} \delta \Pi^{\text{HR}}(\theta, \mathbf{q})[\delta \mathbf{q}] = - \int_{\Omega} [\delta \mathbf{q} \cdot \mathbf{D}^{-1} \mathbf{q}] \, d\Omega \\ \quad - \int_{\Omega} [\delta \mathbf{q} \cdot \nabla \theta] \, d\Omega = 0 \\ \delta \Pi^{\text{HR}}(\theta, \mathbf{q})[\delta \theta] = - \int_{\Omega} [(\nabla \delta \theta) \cdot \mathbf{q}] \, d\Omega \\ \quad + \int_{\Omega} \delta \theta b \, d\Omega = 0 \end{cases} \quad (18)$$

which is now equivalent to the search of a saddle point. Changing sign to both equations and introducing the approximation

$$\begin{cases} \theta \approx \theta^h = N_k^{\theta} \hat{\theta}_k \\ \mathbf{q} \approx \mathbf{q}^h = N_k^{\mathbf{q}} \hat{\mathbf{q}}_k \end{cases} \quad (19)$$

as well as a similar approximation for the corresponding variation fields, equation (18) can be rewritten in matricial form as follows

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\mathbf{q}} \end{Bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{g} \end{Bmatrix} \quad (20)$$

where

$$\begin{cases} \mathbf{A}|_{ij} = \int_{\Omega} [N_i^{\mathbf{q}} \cdot \mathbf{D}^{-1} N_j^{\mathbf{q}}] \, d\Omega, & \hat{\mathbf{q}}|_j = \hat{\mathbf{q}}_j \\ \mathbf{B}|_{ij} = \int_{\Omega} [\nabla N_i^{\theta} \cdot N_j^{\mathbf{q}}] \, d\Omega, & \hat{\boldsymbol{\theta}}|_i = \hat{\theta}_i \\ \mathbf{g}|_i = \int_{\Omega} [N_i^{\theta} b] \, d\Omega \end{cases} \quad (21)$$

Starting from the Hellinger–Reissner functional (17) previously addressed, the following *modified Hellinger–Reissner* functional can be also generated

$$\Pi^{\text{HR,m}}(\theta, \mathbf{q}) = -\frac{1}{2} \int_{\Omega} [\mathbf{q} \cdot \mathbf{D}^{-1} \mathbf{q}] \, d\Omega + \int_{\Omega} [\theta \operatorname{div} \mathbf{q}] \, d\Omega + \int_{\Omega} \theta b \, d\Omega \quad (22)$$

and, requiring its stationarity, we obtain

$$\begin{cases} d\Pi^{\text{HR,m}}(\theta, \mathbf{q})[\delta \mathbf{q}] = -\int_{\Omega} [\delta \mathbf{q} \cdot \mathbf{D}^{-1} \mathbf{q}] \, d\Omega + \int_{\Omega} [\operatorname{div}(\delta \mathbf{q}) \theta] \, d\Omega = 0 \\ d\Pi^{\text{HR,m}}(\theta, \mathbf{q})[\delta \theta] = \int_{\Omega} [\delta \theta \operatorname{div} \mathbf{q}] \, d\Omega + \int_{\Omega} [\delta \theta b] \, d\Omega = 0 \end{cases} \quad (23)$$

which is again equivalent to the search of a saddle point. Changing sign to both equations and introducing again field approximation (19), equation (23) can be rewritten in matrix form as equation (20), with the difference that now

$$\mathbf{B}|_{r,j} = -\int_{\Omega} [N_i^q \operatorname{div}(\mathbf{N}_j^q)] \, d\Omega \quad (24)$$

Similarly, we may also associate an integral form to the three field equation (7) starting from the even more general *Hu–Washizu functional*

$$\Pi^{\text{HW}}(\theta, \mathbf{e}, \mathbf{q}) = \frac{1}{2} \int_{\Omega} [\mathbf{e} \cdot \mathbf{D} \mathbf{e}] \, d\Omega + \int_{\Omega} [\mathbf{q} \cdot (\mathbf{e} - \nabla \theta)] \, d\Omega + \int_{\Omega} \theta b \, d\Omega \quad (25)$$

Requiring the stationarity of functional (25), we obtain

$$\begin{cases} d\Pi^{\text{HW}}(\theta, \mathbf{e}, \mathbf{q})[\delta \mathbf{e}] = \int_{\Omega} [\delta \mathbf{e} \cdot \mathbf{D} \mathbf{e}] \, d\Omega + \int_{\Omega} [\delta \mathbf{e} \cdot \mathbf{q}] \, d\Omega = 0 \\ d\Pi^{\text{HW}}(\theta, \mathbf{e}, \mathbf{q})[\delta \mathbf{q}] = \int_{\Omega} [\delta \mathbf{q} \cdot (\mathbf{e} - \nabla \theta)] \, d\Omega = 0 \\ d\Pi^{\text{HW}}(\theta, \mathbf{e}, \mathbf{q})[\delta \theta] = -\int_{\Omega} [(\nabla \delta \theta) \cdot \mathbf{q}] \, d\Omega + \int_{\Omega} [\delta \theta b] \, d\Omega = 0 \end{cases} \quad (26)$$

which is equivalent to searching a saddle point. Introducing the following approximation

$$\begin{cases} \theta \approx \theta^h = N_k^{\theta} \hat{\theta}_k \\ \mathbf{e} \approx \mathbf{e}^h = N_k^e \hat{\mathbf{e}}_k \\ \mathbf{q} \approx \mathbf{q}^h = N_k^q \hat{\mathbf{q}}_k \end{cases} \quad (27)$$

as well as a similar approximation for the corresponding variation fields, equation (26) can be rewritten in matrix form as follows:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T & \mathbf{0} \\ \mathbf{B} & \mathbf{0} & \mathbf{C}^T \\ \mathbf{0} & \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\theta} \\ \hat{\mathbf{e}} \\ \hat{\mathbf{q}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{h} \end{bmatrix} \quad (28)$$

where

$$\begin{cases} \mathbf{A}|_{ij} = \int_{\Omega} [N_i^{\theta} \cdot \mathbf{D} N_j^{\theta}] \, d\Omega, & \hat{\mathbf{e}}|_i = \hat{\mathbf{e}}_i \\ \mathbf{B}|_{r,j} = \int_{\Omega} [\nabla N_r^{\theta} \cdot \mathbf{N}_j^q] \, d\Omega, & \hat{\mathbf{q}}|_r = \hat{\mathbf{q}}_r \\ \mathbf{C}|_{rs} = -\int_{\Omega} [\nabla N_r^e \cdot \mathbf{N}_s^q] \, d\Omega, & \hat{\theta}|_s = \hat{\theta}_s \\ \mathbf{h}|_s = -\int_{\Omega} [N_s^q b] \, d\Omega \end{cases} \quad (29)$$

For later considerations, we note that equation (28) can be also rewritten as

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{h} \end{bmatrix} \quad (30)$$

where we made the following simple identifications

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix}, & \mathbf{B} &= \{\mathbf{0}, \mathbf{C}\} \\ \mathbf{x} &= \begin{bmatrix} \hat{\theta} \\ \hat{\mathbf{e}} \end{bmatrix}, & \mathbf{y} &= \hat{\mathbf{q}} \end{aligned} \quad (31)$$

Examples of specific choices for the interpolating functions (14), (19), or (27) respectively within the single field, two field, and three field formulations can be found in standard textbooks (Bathe, 1996; Ottosen and Petersson, 1992; Brezzi and Fortin, 1991; Hughes, 1987; Zienkiewicz and Taylor, 2000a) or in the literature.

2.2 Stokes equations

The physical problem

Indicating with \mathbf{u} the fluid velocity, $\boldsymbol{\varepsilon}$ the symmetric part of the velocity gradient, $\boldsymbol{\sigma}$ the stress, p a pressure-like quantity, and with \mathbf{b} the assigned body load per unit volume,

the steady state flow of an incompressible Newtonian fluid can be formulated as a $(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\sigma}, p)$ four field problem as follows:

$$\begin{cases} \operatorname{div} \boldsymbol{\sigma} + \mathbf{b} = \mathbf{0} & \text{in } \Omega \\ \boldsymbol{\sigma} = 2\mu \boldsymbol{\varepsilon} - p \mathbf{1} & \text{in } \Omega \\ \boldsymbol{\varepsilon} = \nabla^s \mathbf{u} & \text{in } \Omega \\ \operatorname{div} \mathbf{u} = 0 & \text{in } \Omega \end{cases} \quad (32)$$

which are respectively the balance, the constitutive, the compatibility, and the incompressibility constraint equations. In particular, ∇^s indicates the symmetric part of the gradient; that is, in a more explicit form,

$$\boldsymbol{\varepsilon} = \nabla^s \mathbf{u} = \frac{1}{2} [\nabla \mathbf{u} + (\nabla \mathbf{u})^T] \quad (33)$$

while the constitutive equation relates the stress $\boldsymbol{\sigma}$ to the symmetric part of the velocity gradient $\boldsymbol{\varepsilon}$ through a material constant μ known as viscosity, and a volumetric pressure-like scalar contribution p .

This set of equations is completed by proper boundary conditions. As for the thermal problem, we prescribe trivial essential conditions on the whole domain boundary, that is,

$$\mathbf{u} = \mathbf{0} \quad \text{on } \partial\Omega \quad (34)$$

As classically done, equation (32) can be simplified eliminating $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$, obtaining a (\mathbf{u}, p) two field problem

$$\begin{cases} \mu \Delta \mathbf{u} - \nabla p + \mathbf{b} = \mathbf{0} & \text{in } \Omega \\ \operatorname{div} \mathbf{u} = 0 & \text{in } \Omega \end{cases} \quad (35)$$

Variational principles

Equation (35) can be derived starting from the *potential energy functional*

$$\Pi(\mathbf{u}) = \frac{1}{2} \mu \int_{\Omega} [\nabla \mathbf{u} : \nabla \mathbf{u}] \, d\Omega - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] \, d\Omega \quad (36)$$

where now \mathbf{u} is a function satisfying the constraint, that is, such that $\operatorname{div} \mathbf{u} = 0$.

To remove the constraint on \mathbf{u} , we can modify the variational principle introducing the functional

$$\begin{aligned} L(\mathbf{u}, p) &= \frac{1}{2} \mu \int_{\Omega} [\nabla \mathbf{u} : \nabla \mathbf{u}] \, d\Omega - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] \, d\Omega \\ &\quad - \int_{\Omega} [p \operatorname{div} \mathbf{u}] \, d\Omega \end{aligned} \quad (37)$$

where p now plays the role of *Lagrange multiplier*.

Requiring the stationarity of functional (37), we obtain

$$\begin{cases} dL(\mathbf{u}, p)[\delta \mathbf{u}] = \mu \int_{\Omega} [(\nabla \delta \mathbf{u}) : \nabla \mathbf{u}] \, d\Omega - \int_{\Omega} [\delta \mathbf{u} \cdot \mathbf{b}] \, d\Omega \\ \quad - \int_{\Omega} [\operatorname{div}(\delta \mathbf{u}) p] \, d\Omega = 0 \\ dL(\mathbf{u}, p)[\delta p] = - \int_{\Omega} [\delta p \operatorname{div} \mathbf{u}] \, d\Omega = 0 \end{cases} \quad (38)$$

which is equivalent to the search of a saddle point. Introducing the following approximation

$$\begin{cases} \mathbf{u} \approx \mathbf{u}^h = N_k^u \hat{\mathbf{u}}_k \\ p \approx p^h = N_k^p \hat{p}_k \end{cases} \quad (39)$$

as well as a similar approximation for the corresponding variation fields, equation (38) can be rewritten as follows

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} \quad (40)$$

where

$$\begin{cases} \mathbf{A}|_{ij} = \mu \int_{\Omega} [\nabla N_i^u : \nabla N_j^u] \, d\Omega, & \hat{\mathbf{u}}|_i = \hat{\mathbf{u}}_i \\ \mathbf{B}|_{r,j} = - \int_{\Omega} [N_r^p \operatorname{div}(\mathbf{N}_j^u)] \, d\Omega, & \hat{p}|_r = \hat{p}_r \\ \mathbf{f}|_i = \int_{\Omega} [N_i^u \cdot \mathbf{b}] \, d\Omega \end{cases} \quad (41)$$

Examples of specific choices for the interpolating functions (39) can be found in standard textbooks (Bathe, 1996; Brezzi and Fortin, 1991; Hughes, 1987; Quarteroni and Valli, 1994; Zienkiewicz and Taylor, 2000a) or in the literature.

2.3 Elasticity

The physical problem

Indicating with \mathbf{u} the body displacement, $\boldsymbol{\varepsilon}$ the strain, $\boldsymbol{\sigma}$ the stress, and with \mathbf{b} the assigned body load per unit volume, the steady state equations for a deformable solid under the assumption of small displacement gradients can be formulated as a $(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\sigma})$ three field problem as follows

$$\begin{cases} \operatorname{div} \boldsymbol{\sigma} + \mathbf{b} = \mathbf{0} & \text{in } \Omega \\ \boldsymbol{\sigma} = \mathbf{D} \boldsymbol{\varepsilon} & \text{in } \Omega \\ \boldsymbol{\varepsilon} = \nabla^s \mathbf{u} & \text{in } \Omega \end{cases} \quad (42)$$

which are respectively the balance, the constitutive, and the compatibility equations.

In particular, we assume a linear constitutive equation, where \mathbb{D} is the elastic material-dependent fourth-order tensor; in the simple case of a mechanically isotropic material, \mathbb{D} specializes as

$$\mathbb{D} = 2\mu\mathbb{I} + \lambda\mathbb{I} \otimes \mathbb{I} \quad (43)$$

and the constitutive equation can be rewritten as

$$\sigma = 2\mu\epsilon + \lambda \operatorname{tr}(\epsilon) \mathbb{I} \quad (44)$$

where $\operatorname{tr}(\epsilon) = \mathbb{I} : \epsilon$. This set of equations is completed by proper boundary conditions. As previously done, we prescribe trivial essential conditions on the whole domain boundary, that is,

$$\mathbf{u} = \mathbf{0} \quad \text{on } \partial\Omega \quad (45)$$

This position is once more very restrictive from a physical point of view but it is still adopted since it simplifies the forthcoming discussion, at the same time without limiting our numerical considerations.

The three field problem (42) can be simplified eliminating the strain ϵ , obtaining a (\mathbf{u}, σ) two field problem

$$\begin{cases} \operatorname{div} \sigma + \mathbf{b} = \mathbf{0} & \text{in } \Omega \\ \sigma = \mathbb{D} \nabla^s \mathbf{u} & \text{in } \Omega \end{cases} \quad (46)$$

and the two field problem (46) can be simplified eliminating the stress σ (or eliminating ϵ and σ directly from equation (42)), obtaining a \mathbf{u} single field problem

$$\operatorname{div}(\mathbb{D} \nabla^s \mathbf{u}) + \mathbf{b} = \mathbf{0} \quad \text{in } \Omega \quad (47)$$

In the case of an isotropic and homogeneous body, this last equation specializes as follows:

$$2\mu \operatorname{div}(\nabla^s \mathbf{u}) + \lambda \nabla(\operatorname{div} \mathbf{u}) + \mathbf{b} = \mathbf{0} \quad \text{in } \Omega \quad (48)$$

Variational principles

The single field equation (47) can be easily derived starting from the potential energy functional

$$\Pi(\mathbf{u}) = \frac{1}{2} \int_{\Omega} [\nabla^s \mathbf{u} : \mathbb{D} \nabla^s \mathbf{u}] \, d\Omega - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] \, d\Omega \quad (49)$$

Requiring the stationarity of potential (49), we obtain

$$\begin{aligned} d\Pi(\mathbf{u})[\delta \mathbf{u}] &= \int_{\Omega} [(\nabla^s \delta \mathbf{u}) : \mathbb{D} \nabla^s \mathbf{u}] \, d\Omega \\ &\quad - \int_{\Omega} [\delta \mathbf{u} \cdot \mathbf{b}] \, d\Omega = 0 \end{aligned} \quad (50)$$

where $\delta \mathbf{u}$ indicates a possible variation of the displacement field \mathbf{u} . Since functional (49) is convex, we may note that

the stationarity requirement is equivalent to a minimization. Recalling the notation introduced in equation (4), we may now introduce an interpolation for the displacement field in the form

$$\mathbf{u} \approx \mathbf{u}^h = \mathbf{N}_i^u \hat{\mathbf{u}}_i \quad (51)$$

as well as a similar approximation for the variation field, such that equation (50) can be rewritten as follows:

$$\mathbf{A} \hat{\mathbf{u}} = \mathbf{f} \quad (52)$$

where

$$\begin{cases} \mathbf{A}_{ij} = \int_{\Omega} [\nabla^s \mathbf{N}_i^u : \mathbb{D} \nabla^s \mathbf{N}_j^u] \, d\Omega, & \hat{\mathbf{u}}_i = \hat{\mathbf{u}}_i \\ \mathbf{f}_i = \int_{\Omega} [\mathbf{N}_i^u \cdot \mathbf{b}] \, d\Omega \end{cases} \quad (53)$$

Besides the integral form (50) associated to the single field equation (47), it is also possible to associate an integral form to the two field equation (46) starting now from the more general Hellinger-Reissner functional

$$\begin{aligned} \Pi^{\text{HR}}(\mathbf{u}, \sigma) &= -\frac{1}{2} \int_{\Omega} [\sigma : \mathbb{D}^{-1} \sigma] \, d\Omega + \int_{\Omega} [\sigma : \nabla \mathbf{u}] \, d\Omega \\ &\quad - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] \, d\Omega \end{aligned} \quad (54)$$

Requiring the stationarity of functional (54), we obtain

$$\begin{cases} d\Pi^{\text{HR}}(\mathbf{u}, \sigma)[\delta \sigma] = -\int_{\Omega} [\delta \sigma : \mathbb{D}^{-1} \sigma] \, d\Omega \\ \quad + \int_{\Omega} [\delta \sigma : \nabla \mathbf{u}] \, d\Omega = 0 \\ d\Pi^{\text{HR}}(\mathbf{u}, \sigma)[\delta \mathbf{u}] = \int_{\Omega} [(\nabla \delta \mathbf{u}) : \sigma] \, d\Omega \\ \quad - \int_{\Omega} [\delta \mathbf{u} \cdot \mathbf{b}] \, d\Omega = 0 \end{cases} \quad (55)$$

which is now equivalent to the search of a saddle point. Changing sign to both equations and introducing the approximation

$$\begin{cases} \mathbf{u} \approx \mathbf{u}^h = \mathbf{N}_i^u \hat{\mathbf{u}}_i \\ \sigma \approx \sigma^h = \mathbf{N}_i^{\sigma} \hat{\sigma}_i \end{cases} \quad (56)$$

as well as a similar approximation for the corresponding variation fields, equation (55) can be rewritten in matricial form as follows

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \hat{\sigma} \\ \hat{\mathbf{u}} \end{Bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{g} \end{Bmatrix} \quad (57)$$

where

$$\begin{cases} \mathbf{A}_{ij} = \int_{\Omega} [\mathbf{N}_i^{\sigma} : \mathbb{D}^{-1} \mathbf{N}_j^{\sigma}] \, d\Omega, & \hat{\sigma}_i = \hat{\sigma}_i \\ \mathbf{B}_{ij} = -\int_{\Omega} [\nabla \mathbf{N}_i^{\sigma} : \mathbf{N}_j^u] \, d\Omega, & \hat{\mathbf{u}}_i = \hat{\mathbf{u}}_i \\ \mathbf{g}_i = -\int_{\Omega} [\mathbf{N}_i^{\sigma} \cdot \mathbf{b}] \, d\Omega \end{cases} \quad (58)$$

Starting from equation (54) the following modified Hellinger-Reissner functional can be also generated

$$\begin{aligned} \Pi^{\text{HR},m}(\mathbf{u}, \sigma) &= -\frac{1}{2} \int_{\Omega} [\sigma : \mathbb{D}^{-1} \sigma] \, d\Omega - \int_{\Omega} [\mathbf{u} \cdot \operatorname{div} \sigma] \, d\Omega \\ &\quad - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] \, d\Omega \end{aligned} \quad (59)$$

and, requiring its stationarity, we obtain

$$\begin{cases} d\Pi^{\text{HR},m}(\mathbf{u}, \sigma)[\delta \sigma] = -\int_{\Omega} [\delta \sigma : \mathbb{D}^{-1} \sigma] \, d\Omega \\ \quad - \int_{\Omega} [\operatorname{div}(\delta \sigma) \cdot \mathbf{u}] \, d\Omega = 0 \\ d\Pi^{\text{HR},m}(\mathbf{u}, \sigma)[\delta \mathbf{u}] = -\int_{\Omega} [\delta \mathbf{u} \cdot \operatorname{div} \sigma] \, d\Omega \\ \quad - \int_{\Omega} [\delta \mathbf{u} \cdot \mathbf{b}] \, d\Omega = 0 \end{cases} \quad (60)$$

which is again equivalent to the search of a saddle point. Changing sign to both equations and introducing again field approximation (56), equation (60) can be rewritten in matricial form as equation (57), with the difference that now

$$\mathbf{B}_{ij} = \int_{\Omega} [\mathbf{N}_i^{\sigma} \cdot \operatorname{div}(\mathbf{N}_j^u)] \, d\Omega \quad (61)$$

Similarly, we may also associate an integral form to three field equation (42) starting from the even more general Hu-Washizu functional

$$\begin{aligned} \Pi^{\text{HW}}(\mathbf{u}, \epsilon, \sigma) &= \frac{1}{2} \int_{\Omega} [\epsilon : \mathbb{D} \epsilon] \, d\Omega - \int_{\Omega} [\sigma : (\epsilon - \nabla^s \mathbf{u})] \, d\Omega \\ &\quad - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] \, d\Omega \end{aligned} \quad (62)$$

Requiring the stationarity of functional (62), we obtain

$$\begin{cases} d\Pi^{\text{HW}}(\mathbf{u}, \epsilon, \sigma)[\delta \epsilon] = \int_{\Omega} [\delta \epsilon : \mathbb{D} \epsilon] \, d\Omega \\ \quad - \int_{\Omega} [\delta \epsilon : \sigma] \, d\Omega = 0 \\ d\Pi^{\text{HW}}(\mathbf{u}, \epsilon, \sigma)[\delta \sigma] = -\int_{\Omega} [\delta \sigma : (\epsilon - \nabla^s \mathbf{u})] \, d\Omega = 0 \\ d\Pi^{\text{HW}}(\mathbf{u}, \epsilon, \sigma)[\delta \mathbf{u}] = \int_{\Omega} [(\nabla^s \delta \mathbf{u}) : \sigma] \, d\Omega \\ \quad - \int_{\Omega} [\delta \mathbf{u} \cdot \mathbf{b}] \, d\Omega = 0 \end{cases} \quad (63)$$

which is again equivalent to the search of a saddle point. Introducing the following approximation

$$\begin{cases} \mathbf{u} \approx \mathbf{u}^h = \mathbf{N}_i^u \hat{\mathbf{u}}_i \\ \epsilon \approx \epsilon^h = \mathbf{N}_i^{\epsilon} \hat{\epsilon}_i \\ \sigma \approx \sigma^h = \mathbf{N}_i^{\sigma} \hat{\sigma}_i \end{cases} \quad (64)$$

as well as a similar approximation for the variation fields, equation (63) can be rewritten as follows

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T & \mathbf{0} \\ \mathbf{B} & \mathbf{0} & \mathbf{C}^T \\ \mathbf{0} & \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \hat{\epsilon} \\ \hat{\sigma} \\ \hat{\mathbf{u}} \end{Bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{h} \end{Bmatrix} \quad (65)$$

where

$$\begin{cases} \mathbf{A}_{ij} = \int_{\Omega} [\mathbf{N}_i^{\epsilon} : \mathbb{D} \mathbf{N}_j^{\epsilon}] \, d\Omega, & \hat{\epsilon}_i = \hat{\epsilon}_i \\ \mathbf{B}_{ij} = -\int_{\Omega} [\mathbf{N}_i^{\epsilon} : \mathbf{N}_j^u] \, d\Omega, & \hat{\mathbf{u}}_i = \hat{\mathbf{u}}_i \\ \mathbf{C}_{ij} = \int_{\Omega} [\nabla^s \mathbf{N}_i^{\sigma} : \mathbf{N}_j^{\epsilon}] \, d\Omega, & \hat{\sigma}_i = \hat{\sigma}_i \\ \mathbf{h}_i = \int_{\Omega} [\mathbf{N}_i^{\sigma} \cdot \mathbf{b}] \, d\Omega \end{cases} \quad (66)$$

For later consideration, we note that equation (65) can be rewritten as

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{x} \\ \mathbf{y} \end{Bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{h} \end{Bmatrix} \quad (67)$$

where we made the following simple identifications

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix}, \quad \mathbf{B} = \{\mathbf{0}, \mathbf{C}\} \\ \mathbf{x} &= \begin{Bmatrix} \hat{\epsilon} \\ \hat{\sigma} \end{Bmatrix}, \quad \mathbf{y} = \hat{\mathbf{u}} \end{aligned} \quad (68)$$

Examples of specific choices for the interpolating functions (51), (56), or (64) respectively within the single field,

the two field, and the three field formulations can be found in standard textbooks (Bathe, 1996; Brezzi and Fortin, 1991; Hughes, 1987; Zienkiewicz and Taylor, 2000a) or in the literature.

Toward incompressible elasticity

It is interesting to observe that the strain ϵ , the stress σ and the symmetric gradient of the displacement $\nabla^s \mathbf{u}$ can be easily decomposed respectively in a deviatoric (traceless) part and a volumetric (trace-related) part. In particular, recalling that we indicate with d the Euclidean space dimension, we may set

$$\begin{cases} \epsilon = \epsilon + \frac{\theta}{d} \mathbf{I} & \text{with } \theta = \text{tr}(\epsilon) \\ \sigma = s + p \mathbf{I} & \text{with } p = \frac{\text{tr}(\sigma)}{d} \\ \nabla^s \mathbf{u} = \bar{\nabla}^s \mathbf{u} + \frac{\text{div } \mathbf{u}}{d} \mathbf{I} & \text{with } \text{div } (\mathbf{u}) = \text{tr}(\nabla^s \mathbf{u}) \end{cases} \quad (69)$$

where θ , p , and $\text{div } (\mathbf{u})$ are the volumetric (trace-related) quantities, while ϵ , s , and $\bar{\nabla}^s \mathbf{u}$ are the deviatoric (or traceless) quantities, that is,

$$\text{tr}(\epsilon) = \text{tr}(s) = \text{tr}(\bar{\nabla}^s \mathbf{u}) = 0 \quad (70)$$

Adopting these deviatoric-volumetric decompositions and limiting the discussion for simplicity of notation to the case of an isotropic material, the three field Hu-Washizu functional (62) can be rewritten as

$$\begin{aligned} \Pi^{\text{HW},m}(\mathbf{u}, \epsilon, \theta, s, p) = & \frac{1}{2} \int_{\Omega} [2\mu \epsilon : \epsilon + k \theta^2] d\Omega \\ & - \int_{\Omega} [s : (\epsilon - \bar{\nabla}^s \mathbf{u})] d\Omega - \int_{\Omega} [p(\theta - \text{div } \mathbf{u})] d\Omega \\ & - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] d\Omega \end{aligned} \quad (71)$$

where we introduce the bulk modulus $k = \lambda + 2\mu/d$. If we now require a strong (pointwise) satisfaction of the deviatoric compatibility condition $\epsilon = \bar{\nabla}^s \mathbf{u}$ (obtained from the stationarity of functional (71) with respect to s) as well as a strong (pointwise) satisfaction of the volumetric constitutive equation $p = k\theta$ (obtained from the stationarity of functional (71) with respect to θ), we end up with the following simpler modified Hellinger-Reissner functional

$$\begin{aligned} \Pi^{\text{HR},m}(\mathbf{u}, p) = & \frac{1}{2} \int_{\Omega} [2\mu \bar{\nabla}^s \mathbf{u} : \bar{\nabla}^s \mathbf{u}] d\Omega - \frac{1}{2} \int_{\Omega} \left[\frac{1}{k} p^2 \right] d\Omega \\ & + \int_{\Omega} [p \text{div } \mathbf{u}] d\Omega - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] d\Omega \end{aligned} \quad (72)$$

It is interesting to observe that taking the variation of functional (72) with respect to p , we obtain the correct relation

between the pressure p and the volumetric component of the displacement gradient, that is,

$$p = k \text{div } \mathbf{u} = \left(\lambda + \frac{2}{d} \mu \right) \text{div } \mathbf{u} \quad (73)$$

For the case of incompressibility ($\lambda \rightarrow \infty$ and $k \rightarrow \infty$), functional (72) reduces to the following form

$$\begin{aligned} \Pi^{\text{HR},m}(\mathbf{u}, p) = & \frac{1}{2} \int_{\Omega} [2\mu \bar{\nabla}^s \mathbf{u} : \bar{\nabla}^s \mathbf{u}] d\Omega + \int_{\Omega} [p \text{div } \mathbf{u}] d\Omega \\ & - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] d\Omega \end{aligned} \quad (74)$$

which resembles the potential energy functional (49) for the case of an isotropic material with the addition of the incompressibility constraint $\text{div } \mathbf{u} = 0$ and with the difference that the quadratic term now involves only the deviatoric part of the symmetric displacement gradient and not the whole symmetric displacement gradient. Requiring the stationarity of functional (74), we obtain

$$\begin{cases} d\Pi^{\text{HR},m}(\mathbf{u}, p)[\delta \mathbf{u}] = \int_{\Omega} [2\mu \bar{\nabla}^s \delta \mathbf{u} : \bar{\nabla}^s \mathbf{u}] d\Omega \\ \quad + \int_{\Omega} [\text{div } (\delta \mathbf{u}) p] d\Omega - \int_{\Omega} [\delta \mathbf{u} \cdot \mathbf{b}] d\Omega = 0 \\ d\Pi^{\text{HR},m}(\mathbf{u}, p)[\delta p] = \int_{\Omega} [\delta p \text{div } \mathbf{u}] d\Omega = 0 \end{cases} \quad (75)$$

Introducing the following approximation

$$\begin{cases} \mathbf{u} \approx \mathbf{u}^h = \mathbf{N}_k^T \hat{\mathbf{u}}_k \\ p \approx p^h = N_k^T \hat{p}_k \end{cases} \quad (76)$$

as well as a similar approximation for the corresponding variation fields, equation (75) can be rewritten as follows

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} \quad (77)$$

where

$$\begin{cases} \mathbf{A}_{ij} = 2\mu \int_{\Omega} [\bar{\nabla} \mathbf{N}_i^T : \bar{\nabla} \mathbf{N}_j^T] d\Omega, & \hat{\mathbf{u}}_i = \hat{\mathbf{u}}_i \\ \mathbf{B}_{ij} = \int_{\Omega} [N_j^T \text{div } (\mathbf{N}_i^T)] d\Omega, & \hat{p}_i = \hat{p}_i \\ \mathbf{f}_i = \int_{\Omega} [\mathbf{N}_i^T \cdot \mathbf{b}] d\Omega \end{cases} \quad (78)$$

It is interesting to observe that this approach may result in an unstable discrete formulation since the volumetric components of the symmetric part of the displacement gradient may not be controlled. Examples of specific choices for the

interpolating functions (76) can be found in standard textbooks (Hughes, 1987; Zienkiewicz and Taylor, 2000a) or in the literature.

A different stable formulation can be easily obtained as in the case of Stokes problem. In particular, we may start from the potential energy functional (49), which for an isotropic material specializes as

$$\begin{aligned} \Pi(\mathbf{u}) = & \frac{1}{2} \int_{\Omega} [2\mu (\nabla^s \mathbf{u} : \nabla^s \mathbf{u}) + \lambda (\text{div } \mathbf{u})^2] d\Omega \\ & - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] d\Omega \end{aligned} \quad (79)$$

Introducing now the pressure-like field $\pi = \lambda \text{div } \mathbf{u}$, we can rewrite functional (79) as

$$\begin{aligned} \Pi^{\pi}(\mathbf{u}, \pi) = & \frac{1}{2} \int_{\Omega} \left[2\mu (\nabla^s \mathbf{u} : \nabla^s \mathbf{u}) - \frac{1}{\lambda} \pi^2 \right] d\Omega \\ & + \int_{\Omega} [\pi \text{div } \mathbf{u}] d\Omega - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] d\Omega \end{aligned} \quad (80)$$

We may note that π is a pressure-like quantity, different, however, from the physical pressure p , previously introduced. In fact, π is the Lagrangian multiplier associated to the incompressibility constraint and it can be related to the physical pressure p recalling relation (73)

$$p = k \text{div } \mathbf{u} = \pi + \frac{2}{d} \mu \text{div } \mathbf{u} \quad (81)$$

For the incompressible case ($\lambda \rightarrow \infty$), functional (80) reduces to the following form:

$$\begin{aligned} \Pi^{\pi}(\mathbf{u}, \pi) = & \frac{1}{2} \int_{\Omega} [2\mu \nabla^s \mathbf{u} : \nabla^s \mathbf{u}] d\Omega \\ & + \int_{\Omega} [\pi \text{div } \mathbf{u}] d\Omega - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] d\Omega \end{aligned} \quad (82)$$

Taking the variation of (82) and introducing the following approximation

$$\begin{cases} \mathbf{u} \approx \mathbf{u}^h = \mathbf{N}_k^T \hat{\mathbf{u}}_k \\ \pi \approx \pi^h = N_k^T \hat{\pi}_k \end{cases} \quad (83)$$

as well as a similar approximation for the corresponding variation fields, we obtain a discrete problem of the following form:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{\pi} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} \quad (84)$$

where

$$\begin{cases} \mathbf{A}_{ij} = 2\mu \int_{\Omega} [\nabla^s \mathbf{N}_i^T : \nabla^s \mathbf{N}_j^T] d\Omega, & \hat{\mathbf{u}}_i = \hat{\mathbf{u}}_i \\ \mathbf{B}_{ij} = \int_{\Omega} [N_j^T \text{div } (\mathbf{N}_i^T)] d\Omega, & \hat{\pi}_i = \hat{\pi}_i \\ \mathbf{f}_i = \int_{\Omega} [\mathbf{N}_i^T \cdot \mathbf{b}] d\Omega \end{cases} \quad (85)$$

It is interesting to observe that, in general, this approach results in a stable discrete formulation since the volumetric components of the symmetric part of the displacement gradient are now controlled. Examples of specific choices for the interpolating functions (83) can be found in standard textbooks (Bathe, 1996; Brezzi and Fortin, 1991; Hughes, 1987; Zienkiewicz and Taylor, 2000a) or in the literature.

Enhanced strain formulation

Starting from the work of Simo and Rifai (1990), recently, a lot of attention has been paid to the so-called *enhanced strain formulation*, which can be variationally deduced for example, from the Hu-Washizu formulation (62). As a first step, the method describes the strain ϵ as the sum of a compatible contribution, $\bar{\nabla}^s \mathbf{u}$, and of an incompatible contribution, $\bar{\epsilon}$, that is,

$$\epsilon = \bar{\nabla}^s \mathbf{u} + \bar{\epsilon} \quad (86)$$

Using this position into the Hu-Washizu formulation (62), we obtain the following functional

$$\begin{aligned} \Pi^{\text{enh}}(\mathbf{u}, \bar{\epsilon}, \sigma) = & \frac{1}{2} \int_{\Omega} [(\bar{\nabla}^s \mathbf{u} + \bar{\epsilon}) : \mathbb{D}(\bar{\nabla}^s \mathbf{u} + \bar{\epsilon})] d\Omega \\ & - \int_{\Omega} [\sigma : \bar{\epsilon}] d\Omega - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] d\Omega \end{aligned} \quad (87)$$

Requiring the stationarity of the functional and introducing the following approximation

$$\begin{cases} \mathbf{u} \approx \mathbf{u}^h = \mathbf{N}_k^T \hat{\mathbf{u}}_k \\ \bar{\epsilon} \approx \bar{\epsilon}^h = \mathbf{N}_{\bar{\epsilon}}^T \hat{\bar{\epsilon}}_{\bar{\epsilon}} \\ \sigma \approx \sigma^h = \mathbf{N}_{\sigma}^T \hat{\sigma}_{\sigma} \end{cases} \quad (88)$$

as well as a similar approximation for the variation fields, we obtain the following discrete problem:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T & \mathbf{0} \\ \mathbf{B} & \mathbf{C} & \mathbf{D}^T \\ \mathbf{0} & \mathbf{D} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{\bar{\epsilon}} \\ \hat{\sigma} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad (89)$$

where

$$\begin{cases} \mathbf{A}|_j = \int_{\Omega} [\nabla^T \mathbf{N}_j^T : \mathbf{D} \nabla^T \mathbf{N}_j^T] d\Omega, & \hat{\mathbf{a}}|_j = \hat{\mathbf{u}}_j \\ \mathbf{B}|_r = \int_{\Omega} [\mathbf{N}_r^T : \mathbf{D} \nabla^T \mathbf{N}_j^T] d\Omega, & \hat{\mathbf{a}}|_r = \hat{\mathbf{e}}_r \\ \mathbf{C}|_{rs} = \int_{\Omega} [\mathbf{N}_r^T : \mathbf{D} \mathbf{N}_s^T] d\Omega, & \hat{\mathbf{c}}|_r = \hat{\mathbf{e}}_r \\ \mathbf{D}|_j = - \int_{\Omega} [\mathbf{N}_j^T : \mathbf{N}_r^T] d\Omega, & \mathbf{f}|_j = \int_{\Omega} [\mathbf{N}_j^T : \mathbf{b}] d\Omega \end{cases} \quad (90)$$

For later consideration, we note that equation (89) can be rewritten as

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} \quad (91)$$

where we made the following simple identifications:

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{C} \end{bmatrix}, \quad \mathbf{B} = \{\mathbf{0}, \mathbf{D}\} \\ \mathbf{x} &= \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{e}} \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{y} = \hat{\mathbf{e}} \end{aligned} \quad (92)$$

Examples of specific choices for the interpolating functions can be found in standard textbooks (Zienkiewicz and Taylor, 2000a) or in the literature.

However, the most widely adopted enhanced strain formulation also requires the incompatible part of the strain to be orthogonal to the stress σ

$$\int_{\Omega} [\sigma : \hat{\mathbf{e}}] d\Omega = 0 \quad (93)$$

If we use conditions (86) and (93) into the Hu–Washizu formulation (62), we obtain the following simplified functional:

$$\begin{aligned} \Pi^{\text{enh}}(\mathbf{u}, \hat{\mathbf{e}}) &= \frac{1}{2} \int_{\Omega} [(\nabla^T \mathbf{u} + \hat{\mathbf{e}}) : \mathbf{D} (\nabla^T \mathbf{u} + \hat{\mathbf{e}})] d\Omega \\ &\quad - \int_{\Omega} [\mathbf{b} \cdot \mathbf{u}] d\Omega \end{aligned} \quad (94)$$

which closely resembles a standard displacement-based incompatible approach. Examples of specific choices for the interpolating functions involved in this simplified enhanced formulation can be found in standard textbooks (Zienkiewicz and Taylor, 2000a) or in the literature.

3 STABILITY OF SADDLE-POINTS IN FINITE DIMENSIONS

3.1 Solvability and stability

The examples discussed in Section 2 clearly show that, after discretization, several formulations typically lead to linear algebraic systems of the general form

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} \quad (95)$$

where \mathbf{A} and \mathbf{B} are respectively an $n \times n$ matrix and an $m \times n$ matrix, while \mathbf{x} and \mathbf{y} are respectively an $n \times 1$ vector and $m \times 1$ vector, as well as \mathbf{f} and \mathbf{g} . Discretizations leading to such a system are often indicated as mixed finite element methods and in the following, we present a simple, algebraic version of the abstract theory that rules most applications of mixed methods.

Our first need is clearly to express in proper form solvability conditions for linear systems of type (95) in terms of the properties of the matrices \mathbf{A} and \mathbf{B} . By solvability we mean that for every right-hand side \mathbf{f} and \mathbf{g} system, (95) has a unique solution. It is well known that this property holds *if and only if* the $(n+m) \times (n+m)$ matrix

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \quad (96)$$

is *nonsingular*, that is, if and only if its determinant is different from zero.

In order to have a good numerical method, however, solvability is not enough. An additional property that we also require is stability. We want to see this property with a little more detail. For a solvable finite-dimensional linear system, we always have continuous dependence of the solution upon the data. This means that there exists a constant c such that for every set of vectors $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$ satisfying (95) we have

$$\|\mathbf{x}\| + \|\mathbf{y}\| \leq c(\|\mathbf{f}\| + \|\mathbf{g}\|) \quad (97)$$

This property implies solvability. Indeed, if we assume that (97) holds for every set of vectors $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$ satisfying (95), then, whenever \mathbf{f} and \mathbf{g} are both zero, \mathbf{x} and \mathbf{y} must also be equal to zero. This is another way of saying that the homogeneous system has only the trivial solution, which implies that the determinant of the matrix (96) is different from zero, and hence the system is solvable.

However, Formula (97) deserves another very important comment. Actually, we did not specify the norms adopted for $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$. We had the right to do so, since in finite

dimension all norms are equivalent. Hence, the change of one norm with another would only result in a change of the numerical value of the constant c , but it would not change the basic fact that such a constant exists. However, in dealing with linear systems resulting from the discretization of a partial differential equation we face a slightly different situation. In fact, if we want to analyze the behaviour of a given method when the meshsize becomes smaller and smaller, we must ideally consider a sequence of linear systems whose dimension increases and approaches infinity when the meshsize tends to zero. As it is well known (and it can be also easily verified), the constants involved in the equivalence of different norms depend on the dimension of the space. For instance, in \mathbb{R}^n , the two norms

$$\|\mathbf{x}\|_1 := \sum_{i=1}^n |x_i| \quad \text{and} \quad \|\mathbf{x}\|_2 := \left(\sum_{i=1}^n |x_i|^2 \right)^{1/2} \quad (98)$$

are indeed equivalent, in the sense that there exist two positive constants c_1 and c_2 such that

$$c_1 \|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2 \leq c_2 \|\mathbf{x}\|_1 \quad (99)$$

for all \mathbf{x} in \mathbb{R}^n . However, it can be rather easily checked that the *best* constants one can choose in (99) are

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n} \|\mathbf{x}\|_2 \quad (100)$$

In particular, the first inequality becomes an equality, for instance, when x_1 is equal to 1 and all the other x_i 's are zero, while the second inequality becomes an equality, for instance, when all the x_i are equal to 1.

When considering a sequence of problems with increasing dimension, we have to take into account that n and m become unbounded. It is then natural to ask if, for a given choice of the norms $\|\mathbf{x}\|$, $\|\mathbf{y}\|$, $\|\mathbf{f}\|$, and $\|\mathbf{g}\|$, it is possible to find a constant c independent of the meshsize (say, h), that is, a constant c that makes (97) hold true for all meshsizes.

However, even if inequality (97) holds with a constant c independent of h , it will not provide a good concept of stability unless the four norms are properly chosen (see Remark 18). This is going to be our next task.

3.2 Assumptions on the norms

We start denoting by $\mathbf{X}, \mathbf{Y}, \mathbf{F}, \mathbf{G}$ respectively the spaces of vectors $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$. Then, we assume what follows.

1. The spaces \mathbf{X} and \mathbf{Y} are equipped with norms $\|\cdot\|_{\mathbf{X}}$ and $\|\cdot\|_{\mathbf{Y}}$ for which the matrices \mathbf{A} and \mathbf{B} satisfy the continuity conditions: *there exist two constants M_a and*

M_b , independent of the meshsize, such that for all \mathbf{x} and \mathbf{z} in \mathbf{X} and for all \mathbf{y} in \mathbf{Y}

$$\mathbf{x}^T \mathbf{A} \mathbf{z} \leq M_a \|\mathbf{x}\|_{\mathbf{X}} \|\mathbf{z}\|_{\mathbf{X}} \quad \text{and} \quad \mathbf{x}^T \mathbf{B}^T \mathbf{y} \leq M_b \|\mathbf{x}\|_{\mathbf{X}} \|\mathbf{y}\|_{\mathbf{Y}} \quad (101)$$

Moreover, we suppose there exist symmetric positive definite matrices $\mathbf{M}^{\mathbf{F}}$ and $\mathbf{M}^{\mathbf{G}}$, respectively of dimensions $n \times n$ and $m \times m$, such that

$$\|\mathbf{x}\|_{\mathbf{X}}^2 = \mathbf{x}^T \mathbf{M}^{\mathbf{F}} \mathbf{x} \quad \forall \mathbf{x} \in \mathbf{X} \quad (102)$$

and

$$\|\mathbf{y}\|_{\mathbf{Y}}^2 = \mathbf{y}^T \mathbf{M}^{\mathbf{G}} \mathbf{y} \quad \forall \mathbf{y} \in \mathbf{Y} \quad (103)$$

2. The spaces \mathbf{F} and \mathbf{G} are equipped with norms $\|\cdot\|_{\mathbf{F}}$ and $\|\cdot\|_{\mathbf{G}}$ defined as the dual norms of $\|\cdot\|_{\mathbf{X}}$ and $\|\cdot\|_{\mathbf{Y}}$, that is,

$$\|\mathbf{f}\|_{\mathbf{F}} := \sup_{\mathbf{x} \in \mathbf{X} \setminus \{0\}} \frac{\mathbf{x}^T \mathbf{f}}{\|\mathbf{x}\|_{\mathbf{X}}} \quad \text{and} \quad \|\mathbf{g}\|_{\mathbf{G}} := \sup_{\mathbf{y} \in \mathbf{Y} \setminus \{0\}} \frac{\mathbf{y}^T \mathbf{g}}{\|\mathbf{y}\|_{\mathbf{Y}}} \quad (104)$$

It is worth noting that

- assumptions (102) to (103) mean that the norms for \mathbf{X} and \mathbf{Y} are both induced by an inner product or, in other words, the norms at hand are *hilbertian* (as it happens in most of the applications);
- for every \mathbf{x} and \mathbf{f} in \mathbb{R}^n and for every \mathbf{y} and \mathbf{g} in \mathbb{R}^m , we have

$$\mathbf{x}^T \mathbf{f} \leq \|\mathbf{x}\|_{\mathbf{X}} \|\mathbf{f}\|_{\mathbf{F}} \quad \text{and} \quad \mathbf{y}^T \mathbf{g} \leq \|\mathbf{y}\|_{\mathbf{Y}} \|\mathbf{g}\|_{\mathbf{G}} \quad (105)$$

- combining the continuity condition (101) on $\|\cdot\|_{\mathbf{X}}$ and $\|\cdot\|_{\mathbf{Y}}$ with the dual norm definition (105), for every $\mathbf{x} \in \mathbf{X}$ and for every $\mathbf{y} \in \mathbf{Y}$, we have the following relations:

$$\|\mathbf{A} \mathbf{x}\|_{\mathbf{F}} = \sup_{\mathbf{z} \in \mathbf{X} \setminus \{0\}} \frac{\mathbf{z}^T \mathbf{A} \mathbf{x}}{\|\mathbf{z}\|_{\mathbf{X}}} \leq M_a \|\mathbf{x}\|_{\mathbf{X}} \quad (106)$$

$$\|\mathbf{B} \mathbf{x}\|_{\mathbf{G}} = \sup_{\mathbf{z} \in \mathbf{Y} \setminus \{0\}} \frac{\mathbf{z}^T \mathbf{B} \mathbf{x}}{\|\mathbf{z}\|_{\mathbf{Y}}} \leq M_b \|\mathbf{x}\|_{\mathbf{X}} \quad (107)$$

$$\|\mathbf{B}^T \mathbf{y}\|_{\mathbf{F}} = \sup_{\mathbf{z} \in \mathbf{X} \setminus \{0\}} \frac{\mathbf{z}^T \mathbf{B}^T \mathbf{y}}{\|\mathbf{z}\|_{\mathbf{X}}} \leq M_b \|\mathbf{y}\|_{\mathbf{Y}} \quad (108)$$

- if \mathbf{A} is symmetric and positive semidefinite, then for every $\mathbf{x}, \mathbf{z} \in \mathbf{X}$

$$|\mathbf{z}^T \mathbf{A} \mathbf{x}| \leq (\mathbf{z}^T \mathbf{A} \mathbf{z})^{1/2} (\mathbf{x}^T \mathbf{A} \mathbf{x})^{1/2} \quad (109)$$

so that (106) can be improved to

$$\|Ax\|_F \leq \sup_{x \in X \setminus \{0\}} \frac{z^T Ax}{\|x\|_X} \leq M_a^{1/2} (x^T Ax)^{1/2} \quad (110)$$

We are now ready to introduce a precise definition of stability.

Stability definition. Given a numerical method, that produces a sequence of matrices A and B when applied to a given sequence of meshes (with the meshsize h going to zero), we choose norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ that satisfy the continuity condition (101), and dual norms $\|\cdot\|_F$ and $\|\cdot\|_G$ according to (104). Then, we say that the method is stable if there exists a constant c , independent of the mesh, such that for all vectors x, y, f, g satisfying the general system (95), it holds

$$\|x\|_X + \|y\|_Y < c(\|f\|_F + \|g\|_G) \quad (111)$$

Having now a precise definition of stability, we can look for suitable assumptions on the matrices A and B that may provide the stability result (111). In particular, to guarantee stability condition (111), we need to introduce two assumptions involving such matrices. The first assumption, the so-called *inf-sup* condition, involves only the matrix B and it will be used throughout the whole section. To illustrate the second assumption we will first focus on a simpler but less general case that involves a 'strong' requirement on the matrix A . Among the problems presented in Section 2, this requirement is verified in practice only for the Stokes problem. Then, we shall tackle a more complex and clearly more general case, corresponding to a 'weak' requirement on the matrix A , suited for instance for discretizations of the mixed formulations of thermal diffusion problems.

Later on we shall deal with some additional complications that occur for instance, in the (u, π) -formulation of nearly incompressible elasticity (cf. (80)). Finally, we shall briefly discuss more complicated problems, omitting the proofs for simplicity.

3.3 A requirement on the B matrix: the *inf-sup* condition

The basic assumption that we are going to use, throughout the whole section, deals with the matrix B . We assume the following:

Inf-sup condition. There exists a positive constant β , independent of the meshsize h , such that:

$$\forall y \in Y \quad \exists x \in X \setminus \{0\} \text{ such that } x^T B^T y \geq \beta \|x\|_X \|y\|_Y \quad (112)$$

Condition (112) requires the existence of a positive constant β , independent of h , such that for every $y \in Y$ we can find a suitable $x \in X$, different from 0 (and depending on y), such that (112) holds.

Remark 1. To better understand the meaning of (112), it might be useful to see when it fails. We thus consider the following $m \times n$ pseudodiagonal matrix ($m < n$)

$$B = \begin{bmatrix} \phi_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \phi_2 & 0 & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & \phi_m & 0 & \dots & 0 \end{bmatrix} \quad (113)$$

with $0 \leq \phi_1 \leq \phi_2 \leq \dots \leq \phi_m \leq 1$. To fix ideas, we suppose that both $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ are equipped with the standard Euclidean norms, which coincide with the corresponding dual norms on F and G (cf. (104)). If $\phi_1 = 0$, choosing $y = (1, 0, \dots, 0)^T \neq 0$, we have $B^T y = 0$. Therefore, for every $x \in X$, we get $x^T B^T y = 0$ and condition (112) cannot hold since β must be positive. We then infer that condition (112) requires that

$$\text{no } y \neq 0 \text{ satisfies } B^T y = 0$$

which, by definition, means that B^T is injective. However, the injectivity of B^T is not sufficient for the fulfillment of condition (112). Indeed, for $0 < \phi_1 \leq \phi_2 \leq \dots \leq \phi_m \leq 1$, the matrix B^T is injective and we have, still choosing $y = (1, 0, \dots, 0)^T$,

$$B^T y = (\phi_1, 0, \dots, 0)^T \neq 0 \quad (114)$$

Since for every $x \in X$ it holds

$$x^T B^T y = \phi_1 x_1 \leq \phi_1 \|x\|_X = \phi_1 \|x\|_X \|y\|_Y \quad (115)$$

we obtain that the constant β in (112) is forced to satisfy

$$0 < \beta \leq \phi_1 \quad (116)$$

As a consequence, if $\phi_1 > 0$ tends to zero with the meshsize h , the matrix B^T is still injective but condition (112) fails, because β , on top of being positive, must be independent of h . Noting that (see (114))

$$\frac{\|B^T y\|_F}{\|y\|_Y} = \phi_1 \quad (117)$$

we then deduce that condition (112) requires that for $y \neq 0$

the vector $B^T y$ is not 'too small' with respect to y

which is a property stronger than the injectivity of the matrix B^T . We will see in Proposition 1 that all these considerations on the particular matrix B in (113) does extend to the general case.

We now rewrite condition (112) in different equivalent forms, which will also make clear the reason why it is called *inf-sup* condition.

Since, by assumption, x is different from zero, condition (112) can equivalently be written as

$$\forall y \in Y \quad \exists x \in X \setminus \{0\} \text{ such that } \frac{x^T B^T y}{\|x\|_X} \geq \beta \|y\|_Y \quad (118)$$

This last form (118) highlights that given $y \in Y$, the most suitable $x \in X$ is the one that makes the left-hand side of (118) as big as possible. Hence, the best we can do is to take the *supremum* of the left-hand side, when x varies among all possible $x \in X$ different from 0. Hence, we may equivalently require that

$$\forall y \in Y \quad \sup_{x \in X \setminus \{0\}} \frac{x^T B^T y}{\|x\|_X} \geq \beta \|y\|_Y \quad (119)$$

In a sense, we got rid of the task of choosing x . However, condition (119) still depends on y and it clearly holds for $y = 0$. Therefore, we can concentrate on the y 's that are different from 0; in particular, for $y \neq 0$ condition (119) can be also written as

$$\sup_{x \in X \setminus \{0\}} \frac{x^T B^T y}{\|x\|_X \|y\|_Y} \geq \beta \quad (120)$$

The worst possible y is therefore the one that makes the left-hand side of (120) as small as possible. If we want (120) to hold for every $y \in Y$ we might as well consider the worst case, looking directly at the *infimum* of the left-hand side of (120) among all possible y 's, requiring that

$$\inf_{y \in Y \setminus \{0\}} \sup_{x \in X \setminus \{0\}} \frac{x^T B^T y}{\|x\|_X \|y\|_Y} \geq \beta \quad (121)$$

The advantage of formulation (121), if any, is that we got rid of the dependency on y as well. Indeed, condition (121) is now a condition on the matrix B , on the spaces X and Y (together with their norms) as well as on the crucial constant β .

Let us see now the relationship of the *inf-sup* condition with a basic property of the matrix B .

Proposition 1. The *inf-sup* condition (112) is equivalent to require that

$$\beta \|y\|_Y \leq \|B^T y\|_F \quad \forall y \in Y \quad (122)$$

Therefore, in particular, the *inf-sup* condition implies that the matrix B^T is injective.

Proof. Assume that the *inf-sup* condition (112) holds, and let y be any vector in Y . By the equivalent form (119) and using definition (104) of the dual norm $\|\cdot\|_F$, we have

$$\beta \|y\|_Y \leq \sup_{x \in X \setminus \{0\}} \frac{x^T B^T y}{\|x\|_X} = \|B^T y\|_F \quad (123)$$

and therefore (122) holds true. Moreover, the matrix B^T is injective since (122) shows that $y \neq 0$ implies $B^T y \neq 0$.

Assume conversely that (122) holds. Using again the definition (104) of the dual norm $\|\cdot\|_F$, we have

$$\beta \|y\|_Y \leq \|B^T y\|_F = \sup_{x \in X \setminus \{0\}} \frac{x^T B^T y}{\|x\|_X} \quad (124)$$

which implies the *inf-sup* condition in the form (119). \square

Remark 2. Whenever the $m \times n$ matrix B satisfies the *inf-sup* condition, the injectivity of B^T implies that $n \geq m$. We point out once again (cf. Remark 1) that the injectivity of B^T is not sufficient for the fulfillment of the *inf-sup* condition.

Additional relationships between the *inf-sup* and other properties of the matrix B will be presented later on in Section 3.5.

3.4 A 'strong' condition on the A matrix. Ellipticity on the whole space — Stokes

As we shall see in the sequel, the *inf-sup* condition is a necessary condition for having stability of problems of the general form (95). In order to have sufficient conditions, we now introduce a further assumption on the matrix A . As discussed at the end of Section 3.2, we start considering a strong condition on the matrix A . More precisely, we assume the following:

Ellipticity condition. There exists a positive constant α , independent of the meshsize h , such that

$$\alpha \|x\|_X^2 \leq x^T A x \quad \forall x \in X \quad (125)$$

We first notice that from (101) and (125) it follows that

$$\alpha \leq M_a \quad (126)$$

We have now the following theorem.

Theorem 1. Let x, y, f, g satisfy the general system of equations (95). Moreover, assume that A is symmetric and

that the continuity conditions (101), the dual norm assumptions (105), the inf-sup (112) and the ellipticity requirement (125) are all satisfied. Then, we have

$$\|\mathbf{x}\|_X \leq \frac{1}{\alpha} \|\mathbf{f}\|_F + \frac{M_a^{1/2}}{\alpha^{1/2}\beta} \|\mathbf{g}\|_G \quad (127)$$

$$\|\mathbf{y}\|_Y \leq \frac{2M_a^{1/2}}{\alpha^{1/2}\beta} \|\mathbf{f}\|_F + \frac{M_g}{\beta^2} \|\mathbf{g}\|_G \quad (128)$$

Proof. We shall prove the result by splitting $\mathbf{x} = \mathbf{x}_f + \mathbf{x}_g$ and $\mathbf{y} = \mathbf{y}_f + \mathbf{y}_g$ defined as the solutions of

$$\begin{cases} \mathbf{A}\mathbf{x}_f + \mathbf{B}^T\mathbf{y}_f = \mathbf{f} \\ \mathbf{B}\mathbf{x}_f = 0 \end{cases} \quad (129)$$

and

$$\begin{cases} \mathbf{A}\mathbf{x}_g + \mathbf{B}^T\mathbf{y}_g = 0 \\ \mathbf{B}\mathbf{x}_g = \mathbf{g} \end{cases} \quad (130)$$

We proceed in several steps.

• *Step 1 – Estimate of \mathbf{x}_f and $\mathbf{A}\mathbf{x}_f$.* We multiply the first equation of (129) to the left by \mathbf{x}_f^T and we notice that $\mathbf{x}_f^T \mathbf{B}^T \mathbf{y}_f = \mathbf{y}_f^T \mathbf{B} \mathbf{x}_f = 0$ (by the second equation). Hence,

$$\mathbf{x}_f^T \mathbf{A} \mathbf{x}_f = \mathbf{x}_f^T \mathbf{f} \quad (131)$$

and, using the ellipticity condition (125), relation (131), and the first of the dual norm estimates (105), we have

$$\alpha \|\mathbf{x}_f\|_X^2 \leq \mathbf{x}_f^T \mathbf{A} \mathbf{x}_f = \mathbf{x}_f^T \mathbf{f} \leq \|\mathbf{x}_f\|_X \|\mathbf{f}\|_F \quad (132)$$

giving immediately

$$\|\mathbf{x}_f\|_X \leq \frac{1}{\alpha} \|\mathbf{f}\|_F \quad (133)$$

and

$$\mathbf{x}_f^T \mathbf{A} \mathbf{x}_f \leq \frac{1}{\alpha} \|\mathbf{f}\|_F^2 \quad (134)$$

Therefore, using (110) we also get

$$\|\mathbf{A}\mathbf{x}_f\|_F \leq \frac{M_a^{1/2}}{\alpha^{1/2}} \|\mathbf{f}\|_F \quad (135)$$

• *Step 2 – Estimate of \mathbf{y}_f .* We use now the inf-sup condition (112) with $\mathbf{y} = \mathbf{y}_f$. We obtain that there exists $\tilde{\mathbf{x}} \in \mathbf{X}$ such that $\tilde{\mathbf{x}}^T \mathbf{B}^T \mathbf{y}_f \geq \beta \|\tilde{\mathbf{x}}\|_X \|\mathbf{y}_f\|_Y$. Multiplying the first equation of (129) by $\tilde{\mathbf{x}}^T$ and using the first of the dual norm estimates (105), we have

$$\begin{aligned} \beta \|\tilde{\mathbf{x}}\|_X \|\mathbf{y}_f\|_Y &\leq \tilde{\mathbf{x}}^T \mathbf{B}^T \mathbf{y}_f = \tilde{\mathbf{x}}^T (\mathbf{f} - \mathbf{A}\mathbf{x}_f) \\ &\leq \|\tilde{\mathbf{x}}\|_X \|\mathbf{f} - \mathbf{A}\mathbf{x}_f\|_F \end{aligned} \quad (136)$$

We now use the fact that in the inf-sup condition (112) we had $\tilde{\mathbf{x}} \neq 0$, so that in the above equation (136) we can simplify by its norm. Then, using (135) and (126), we obtain

$$\begin{aligned} \|\mathbf{y}_f\|_Y &\leq \frac{1}{\beta} \|\mathbf{f} - \mathbf{A}\mathbf{x}_f\|_F \leq \left(\frac{1}{\beta} + \frac{M_a^{1/2}}{\alpha^{1/2}\beta} \right) \|\mathbf{f}\|_F \\ &\leq \frac{2M_a^{1/2}}{\alpha^{1/2}\beta} \|\mathbf{f}\|_F \end{aligned} \quad (137)$$

• *Step 3 – Estimate of $\mathbf{x}_g^T \mathbf{A} \mathbf{x}_g$ by $\|\mathbf{y}_g\|_Y$.* We multiply the first equation of (130) by \mathbf{x}_g^T . Using the second equation of (130) and the second of the dual norm estimates (105), we have

$$\mathbf{x}_g^T \mathbf{A} \mathbf{x}_g = -\mathbf{x}_g^T \mathbf{B}^T \mathbf{y}_g = \mathbf{y}_g^T \mathbf{B} \mathbf{x}_g = \mathbf{y}_g^T \mathbf{g} \leq \|\mathbf{y}_g\|_Y \|\mathbf{g}\|_G \quad (138)$$

• *Step 4 – Estimate of $\|\mathbf{y}_g\|_Y$ by $(\mathbf{x}_g^T \mathbf{A} \mathbf{x}_g)^{1/2}$.* We proceed as in Step 2. Using the inf-sup condition (112) with $\mathbf{y} = \mathbf{y}_g$ we get a new vector, that we call again $\tilde{\mathbf{x}}$, such that $\tilde{\mathbf{x}}^T \mathbf{B}^T \mathbf{y}_g \geq \beta \|\tilde{\mathbf{x}}\|_X \|\mathbf{y}_g\|_Y$. This relation, the first equation of (130), and the continuity property (109) yield

$$\begin{aligned} \beta \|\tilde{\mathbf{x}}\|_X \|\mathbf{y}_g\|_Y &\leq \tilde{\mathbf{x}}^T \mathbf{B}^T \mathbf{y}_g = -\tilde{\mathbf{x}}^T \mathbf{A} \mathbf{x}_g \\ &\leq M_a^{1/2} \|\tilde{\mathbf{x}}\|_X (\mathbf{x}_g^T \mathbf{A} \mathbf{x}_g)^{1/2} \end{aligned} \quad (139)$$

giving

$$\|\mathbf{y}_g\|_Y \leq \frac{M_a^{1/2}}{\beta} (\mathbf{x}_g^T \mathbf{A} \mathbf{x}_g)^{1/2} \quad (140)$$

• *Step 5 – Estimate of $\|\mathbf{x}_g\|_X$ and $\|\mathbf{y}_g\|_Y$.* We first combine (138) and (140) to obtain

$$\|\mathbf{y}_g\|_Y \leq \frac{M_g}{\beta^2} \|\mathbf{g}\|_G \quad (141)$$

Moreover, using the ellipticity assumption (125) in (138) and inserting (141), we have

$$\alpha \|\mathbf{x}_g\|_X^2 \leq \mathbf{x}_g^T \mathbf{A} \mathbf{x}_g \leq \|\mathbf{y}_g\|_Y \|\mathbf{g}\|_G \leq \frac{M_g}{\beta^2} \|\mathbf{g}\|_G^2 \quad (142)$$

which can be rewritten as

$$\|\mathbf{x}_g\|_X \leq \frac{M_g^{1/2}}{\alpha^{1/2}\beta} \|\mathbf{g}\|_G \quad (143)$$

The final estimate follows then by simply collecting the separate estimates (133), (137), (143), and (141). □

A straightforward consequence of Theorem 1 and Remark 4 is the following stability result (cf. (111)):

Corollary 1. Assume that a numerical method produces a sequence of matrices \mathbf{A} and \mathbf{B} for which both the inf-sup condition (112) and the ellipticity condition (125) are satisfied. Then the method is stable.

Remark 3. In certain applications, it might happen that the constants α and β either depend on h (and tend to zero as h tends to zero) or have a fixed value that is however very small. It is therefore important to keep track of the possible degeneracy of the constants in our estimates when α and/or β are very small. In particular, it is relevant to know whether our stability constants degenerate, say, as $1/\beta$, or $1/\beta^2$, or other powers of $1/\beta$ (and, similarly, of $1/\alpha$). In this respect, we point out that the behavior indicated in (127) and (128) is optimal. This means that we cannot hope to find a better proof giving a better behavior of the constants in terms of powers of $1/\alpha$ and $1/\beta$. Indeed, consider the system

$$\begin{bmatrix} 2 & \sqrt{a} & b \\ \sqrt{a} & a & 0 \\ b & 0 & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ y \end{Bmatrix} = \begin{Bmatrix} f_1 \\ f_2 \\ g \end{Bmatrix} \quad 0 < a, b \ll 1 \quad (144)$$

whose solution is

$$x_1 = \frac{g}{b}, \quad x_2 = \frac{f_2}{a} - \frac{g}{a^{1/2}b}, \quad y = \frac{f_1}{b} - \frac{f_2}{a^{1/2}b} - \frac{g}{b^2} \quad (145)$$

Since the constants α and β are given by

$$\alpha = \frac{2 + a - \sqrt{a^2 + 4}}{2} = \frac{4a}{2(2 + a + \sqrt{a^2 + 4})} \approx a$$

and

$$\beta = b$$

we see from (145) that there are cases in which the actual stability constants behave exactly as predicted by the theory.

Remark 4. We point out that the symmetry condition on the matrix \mathbf{A} is not necessary. Indeed, with a slightly different (and even simpler) proof one can prove stability without the symmetry assumption. The dependence of the stability constant upon α and β is however worse, as it can be seen in the following example. Considering the system

$$\begin{bmatrix} 1 & -1 & b \\ 1 & a & 0 \\ b & 0 & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ y \end{Bmatrix} = \begin{Bmatrix} f_1 \\ f_2 \\ g \end{Bmatrix} \quad 0 < a, b \ll 1 \quad (146)$$

one easily obtains

$$x_1 = \frac{g}{b}, \quad x_2 = \frac{f_2}{a} - \frac{g}{ab}, \quad y = \frac{f_1}{b} + \frac{f_2}{ab} - \frac{(1+a)g}{ab^2} \quad (147)$$

Since $\alpha = a$ and $\beta = b$, from (147) we deduce that the bounds of Theorem 1 cannot hold when \mathbf{A} is not symmetric.

As announced in the title of the section the situation in which \mathbf{A} is elliptic in the whole space is typical (among others) of the Stokes problem, as presented in (22) to (24). Indeed, denoting the interpolating functions for \mathbf{u} and p by \mathbf{N}_h^u and N_h^p respectively (cf. (39)), if we set

$$\|\mathbf{u}\|_X^2 := \mu \int_{\Omega} |\nabla(\mathbf{N}_h^u \hat{\mathbf{u}})|^2 \, d\Omega \quad (148)$$

and

$$\|\hat{\mathbf{u}}\|_Y^2 := \int_{\Omega} |\nabla_p \hat{p}_r|^2 \, d\Omega \quad (149)$$

we can easily see that conditions (101) are verified with $M_a = 1$ and $M_g = \sqrt{(d/\mu)}$ respectively. Clearly the ellipticity property (125) is also verified with $\alpha = 1$, no matter what is the choice of the mesh and of the interpolating functions. On the other hand, the inf-sup Property (112) is much less obvious, as we are going to see in Section 4, and finite element choices have to be specially tailored in order to satisfy it.

3.5 The inf-sup condition and the lifting operator

In this section, we shall see that the inf-sup condition is related to another important property of the matrix \mathbf{B} . Before proceeding, we recall that an $m \times n$ matrix \mathbf{B} is surjective if for every $\mathbf{g} \in \mathbb{R}^n$, there exists $\mathbf{x}_g \in \mathbb{R}^m$ such that $\mathbf{B}\mathbf{x}_g = \mathbf{g}$.

We have the following Proposition.

Proposition 2. The inf-sup condition (112) is equivalent to require the existence of a lifting operator $\mathbf{L}: \mathbf{g} \rightarrow \mathbf{x}_g = \mathbf{L}\mathbf{g}$ such that, for every $\mathbf{g} \in \mathbb{R}^n$, it holds

$$\begin{cases} \mathbf{B}\mathbf{x}_g = \mathbf{g} \\ \beta \|\mathbf{x}_g\|_X = \beta \|\mathbf{L}\mathbf{g}\|_X \leq \|\mathbf{g}\|_G = \|\mathbf{B}\mathbf{x}_g\|_G \end{cases} \quad (150)$$

Therefore, in particular, the inf-sup condition implies that the matrix \mathbf{B} is surjective.

Proof. We begin by recalling that there exists a symmetric $(n \times n)$ positive definite matrix \mathbf{M}^2 such that (cf. (102))

$$\mathbf{x}^T \mathbf{M}^2 \mathbf{x} = \|\mathbf{x}\|_X^2 \quad (151)$$

It is clear that the choice $\mathbf{A} = \mathbf{M}^T$ easily satisfies the first of the continuity conditions (101) with $M_0 = 1$, as well as the ellipticity condition (125) with $\alpha = 1$. Given \mathbf{g} , if the *inf-sup* condition holds, we can, therefore, use Theorem 1 and find a unique solution $(\tilde{\mathbf{x}}_g, \tilde{\mathbf{y}}_g)$ of the following auxiliary problem:

$$\begin{cases} \mathbf{M}^T \tilde{\mathbf{x}}_g + \mathbf{B}^T \tilde{\mathbf{y}}_g = 0, \\ \mathbf{B} \tilde{\mathbf{x}}_g = \mathbf{g} \end{cases} \quad (152)$$

We can now use estimate (143) from Step 5 of the proof of Theorem 1, recalling that in our case, $\alpha = M_0 = 1$ since we are using the matrix \mathbf{M}^T instead of \mathbf{A} . We obtain

$$\beta \|\tilde{\mathbf{x}}_g\|_X \leq \|\mathbf{g}\|_G \quad (153)$$

It is then clear that setting $\mathbf{Lg} := \tilde{\mathbf{x}}_g$ (the first part of the solution of the auxiliary problem (152)) we have that estimate (150) in our statement holds true.

Assume conversely that we have the existence of a continuous lifting \mathbf{L} satisfying (150). First we recall that there exists a symmetric $(m \times m)$ positive definite matrix \mathbf{M}^T such that (cf. (103))

$$\mathbf{y}^T \mathbf{M}^T \mathbf{y} = \|\mathbf{y}\|_Y^2 \quad (154)$$

Then, for a given $\mathbf{y} \in \mathbf{Y}$, we set first $\mathbf{g} := \mathbf{M}^T \mathbf{y}$ (so that $\mathbf{y}^T \mathbf{g} = \|\mathbf{y}\|_Y^2$) and then we define $\mathbf{x}_g := \mathbf{Lg}$ (so that $\mathbf{Bx}_g = \mathbf{g}$). Hence,

$$\mathbf{x}_g^T \mathbf{B}^T \mathbf{y} = \mathbf{y}^T \mathbf{Bx}_g = \mathbf{y}^T \mathbf{g} = \mathbf{y}^T \mathbf{M}^T \mathbf{y} = \|\mathbf{y}\|_Y^2 \quad (155)$$

On the other hand, it is easy to see that using (150) we have

$$\beta \|\mathbf{x}_g\|_X \leq \|\mathbf{g}\|_G \leq \|\mathbf{y}\|_Y \quad (156)$$

where the last inequality is based on the choice $\mathbf{g} = \mathbf{M}^T \mathbf{y}$ and the use of (108) with \mathbf{M}^T in the place of \mathbf{B}^T . Hence, for every $\mathbf{y} \in \mathbf{Y}$, different from zero, we constructed $\mathbf{x} = \mathbf{x}_g \in \mathbf{X}$, different from zero, which, joining (155) and (156), satisfies

$$\mathbf{x}_g^T \mathbf{B}^T \mathbf{y} = \|\mathbf{y}\|_Y^2 \geq \beta \|\mathbf{x}_g\|_X \|\mathbf{y}\|_Y \quad (157)$$

that is, the *inf-sup* condition in its original form (112). \square

3.6 A 'weak' condition on the \mathbf{A} matrix. Ellipticity on the kernel — thermal diffusion

We now consider that, together with the *inf-sup* condition on \mathbf{B} , the condition on \mathbf{A} is weaker than the full Ellipticity (125). In particular, we require the ellipticity of \mathbf{A} to

hold only in a subspace \mathbf{X}_0 of the whole space \mathbf{X} , with \mathbf{X}_0 defined as follows:

$$\mathbf{X}_0 := \text{Ker}(\mathbf{B}) = \{\mathbf{x} \in \mathbf{X} \text{ such that } \mathbf{Bx} = 0\} \quad (158)$$

More precisely, we require the following:

Elker condition. There exists a positive constant α_0 , independent of the meshsize h , such that

$$\alpha_0 \|\mathbf{x}\|_X^2 \leq \mathbf{x}^T \mathbf{Ax} \quad \forall \mathbf{x} \in \mathbf{X}_0 \quad (159)$$

The above condition is often called *elker* since it requires the ellipticity on the kernel. Moreover, from (101) and (159), we get

$$\alpha_0 \leq M_0 \quad (160)$$

The following theorem generalizes Theorem 1. For the sake of completeness, we present here the proof in the case of a matrix \mathbf{A} that is not necessarily symmetric.

Theorem 2. Let $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ satisfy system (1) and assume that the continuity conditions (101), the dual norm assumptions (105), the *inf-sup* (112), and the *elker* condition (159) are satisfied. Then, we have

$$\|\mathbf{x}\|_X \leq \frac{1}{\alpha_0} \|\mathbf{f}\|_F + \frac{2M_0}{\alpha_0 \beta} \|\mathbf{g}\|_G \quad (161)$$

$$\|\mathbf{y}\|_Y \leq \frac{2M_0}{\alpha_0 \beta} \|\mathbf{f}\|_F + \frac{2M_0^2}{\alpha_0 \beta^2} \|\mathbf{g}\|_G \quad (162)$$

Proof. We first set $\mathbf{x}_g := \mathbf{Lg}$ where \mathbf{L} is the lifting operator defined by Proposition 2. We also point out the following estimates on \mathbf{x}_g : from the continuity of the lifting \mathbf{L} (150), we have

$$\beta \|\mathbf{x}_g\|_X \leq \|\mathbf{g}\|_G \quad (163)$$

and using (106) and (163), we obtain

$$\|\mathbf{Ax}_g\|_F \leq M_0 \|\mathbf{x}_g\|_X \leq \frac{M_0}{\beta} \|\mathbf{g}\|_G \quad (164)$$

Then, we set

$$\mathbf{x}_0 := \mathbf{x} - \mathbf{x}_g = \mathbf{x} - \mathbf{Lg} \quad (165)$$

and we notice that $\mathbf{x}_0 \in \mathbf{X}_0$. Moreover, $(\mathbf{x}_0, \mathbf{y})$ solves the linear system

$$\begin{cases} \mathbf{Ax}_0 + \mathbf{B}^T \mathbf{y} = \mathbf{f} - \mathbf{Ax}_g, \\ \mathbf{Bx}_0 = 0 \end{cases} \quad (166)$$

We can now proceed as in Steps 1 and 2 of the proof of Theorem 1 (as far as we do not use (110), since we gave

up the symmetry assumption). We note that our weaker assumption *elker* (159) is sufficient for allowing the first step in (132). Proceeding as in the first part of Step 1, and using (164) at the end, we get

$$\|\mathbf{x}_0\|_X \leq \frac{1}{\alpha_0} \|\mathbf{f} - \mathbf{Ax}_g\|_F \leq \frac{1}{\alpha_0} \left(\|\mathbf{f}\|_F + \frac{M_0}{\beta} \|\mathbf{g}\|_G \right) \quad (167)$$

This allows to reconstruct the estimate on \mathbf{x} :

$$\begin{aligned} \|\mathbf{x}\|_X &= \|\mathbf{x}_0 + \mathbf{x}_g\|_X \leq \frac{1}{\alpha_0} \|\mathbf{f}\|_F + \left(\frac{M_0}{\alpha_0 \beta} + \frac{1}{\beta} \right) \|\mathbf{g}\|_G \\ &\leq \frac{1}{\alpha_0} \|\mathbf{f}\|_F + \frac{2M_0}{\alpha_0 \beta} \|\mathbf{g}\|_G \end{aligned} \quad (168)$$

where we have used (160) in the last inequality. Combining (106) and (168), we also have

$$\|\mathbf{Ax}\|_F \leq M_0 \|\mathbf{x}\|_X \leq \frac{M_0}{\alpha_0} \|\mathbf{f}\|_F + \frac{2M_0^2}{\alpha_0 \beta} \|\mathbf{g}\|_G \quad (169)$$

which is weaker than (135) since we could not use the symmetry assumption. Then, we proceed as in Step 2 to obtain, as in (137)

$$\beta \|\mathbf{y}\|_Y \leq \|\mathbf{f} - \mathbf{Ax}\|_F \quad (170)$$

and using the above estimate (169) on \mathbf{Ax} in (170), we obtain

$$\begin{aligned} \|\mathbf{y}\|_Y &\leq \left(\frac{1}{\beta} + \frac{M_0}{\alpha_0 \beta} \right) \|\mathbf{f}\|_F + \frac{2M_0^2}{\alpha_0 \beta^2} \|\mathbf{g}\|_G \\ &\leq \frac{2M_0}{\alpha_0 \beta} \|\mathbf{f}\|_F + \frac{2M_0^2}{\alpha_0 \beta^2} \|\mathbf{g}\|_G \end{aligned} \quad (171)$$

and the proof is concluded. \square

A straightforward consequence of Theorem 2 is the following stability result (cf. (111)):

Corollary 2. Assume that a numerical method produces a sequence of matrices \mathbf{A} and \mathbf{B} for which both the *inf-sup* condition (112) and the *elker* condition (159) are satisfied. Then the method is stable.

Remark 5. In the spirit of Remark 3, we notice that the dependence of the stability constants from α_0 and β is optimal, as shown by the previous example (146), for which $\alpha_0 = a$ and $\beta = b$. It is interesting to notice that just adding the assumption that \mathbf{A} is symmetric will not improve the bounds. Indeed, considering the system

$$\begin{bmatrix} 1 & 1 & b \\ 1 & a & 0 \\ b & 0 & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ y \end{Bmatrix} = \begin{Bmatrix} f_1 \\ f_2 \\ g \end{Bmatrix} \quad 0 < a, b \ll 1 \quad (172)$$

one easily obtains

$$x_1 = \frac{g}{b}, \quad x_2 = \frac{f_2}{a} - \frac{g}{ab}, \quad y = \frac{f_1}{b} - \frac{f_2}{ab} - \frac{(1-a)g}{ab^2} \quad (173)$$

Since $\alpha_0 = a$ and $\beta = b$, system (172) shows the same behavior as the bounds of Theorem 2 (and not better), even though \mathbf{A} is symmetric. In order to get back the better bounds found in Theorem 1, we have to assume that \mathbf{A} , on top of satisfying the ellipticity in the kernel (159), is symmetric and positive semidefinite in the whole \mathbb{R}^n (a property that the matrix \mathbf{A} in (172) does not have for $0 \leq a < 1$). This is because, in order to improve the bounds, one has to use (110) that, indeed, requires \mathbf{A} to be symmetric and positive semidefinite.

As announced in the title of the section, the situation in which \mathbf{A} is elliptic only in the kernel of \mathbf{B} is typical (among others) of the mixed formulation of thermal problems, as presented in (22) to (24). As in (19), we denote the interpolating functions for θ and \mathbf{q} by N_θ^h and $N_\mathbf{q}^h$, respectively, and we set

$$\begin{aligned} \|\hat{\mathbf{q}}\|_X^2 &:= \int_\Omega \left[(N_\theta^h \hat{q}_i) \cdot \mathbf{D}^{-1} (N_\mathbf{q}^h \hat{q}_i) \right] d\Omega \\ &\quad + \frac{\ell^2}{k_*} \int_\Omega |\text{div} (N_\mathbf{q}^h \hat{q}_i)|^2 d\Omega \end{aligned} \quad (174)$$

$$\|\hat{\theta}\|_Y^2 := \int_\Omega |N_\theta^h \hat{\theta}|^2 d\Omega \quad (175)$$

where ℓ represents some characteristic length of the domain Ω (for instance its diameter) and k_* represent some characteristic value of the thermal conductivity (for instance, its average).

We can easily see that the continuity conditions (101) are verified with $M_0 = 1$ and $M_1 = \ell^{-1}/k_*$, respectively. On the other hand, the full ellipticity property (125) is verified only with a constant α that behaves, in most cases, like $\alpha \simeq h^2$, where h is a measure of the mesh size. Indeed, the norm of $\hat{\mathbf{q}}$ contains the derivatives of the interpolating functions, while the term $\hat{\mathbf{q}}^T \mathbf{A} \hat{\mathbf{q}}$ does not, as it can be seen in (21). On the other hand, we are obliged to add the divergence term in the definition (174) of the norm of $\hat{\mathbf{q}}$; otherwise, we cannot have a uniform bound for M_0 when the meshsize goes to zero, precisely for the same reason as before. Indeed, the term $\hat{\theta}^T \mathbf{B} \hat{\mathbf{q}}$ contains the derivatives of the interpolating functions N_θ^h (see (21)), and the first part of $\|\hat{\mathbf{q}}\|_X$ does not. One can object that the constant M_0 does not show up in the stability estimates. It does, however, come into play in the error estimates, as we are going to see in Section 5.

It follows from this analysis that, keeping the norms as in (174) and (175), the *elker* property (159) holds, in

practical cases, only if the kernel of \mathbf{B} is made of free-divergence vectors. In that case, we would actually have $\alpha_0 = 1$, no matter what is the choice of the mesh and of the interpolating functions.

On the other hand, the *inf-sup* property (112) is still difficult and it depends heavily on the choices of the interpolating functions. As we are going to see in the next section, the need to satisfy both the *elker* and the *inf-sup* condition poses serious limitations on the choice of the approximations. Apart from some special one-dimensional cases, there is no hope that these two properties can hold at the same time unless the finite element spaces have been *designed* for that. However, this work has been already done and there are several families of finite element spaces that can be profitably used for these problems. We also note that the *elker* condition (or, more precisely, the requirement that the kernel of \mathbf{B} is made only of free-divergence vectors) poses some difficulties in the choice of the element, but in most applications it constitutes a very desirable conservation property for the discrete solutions.

3.7 Perturbation of the problem — nearly incompressible elasticity

We now consider a possible variant of our general form (95). Namely, we assume that we have, together with the matrices \mathbf{A} and \mathbf{B} , a third matrix \mathbf{C} , that we assume to be an $(m \times m)$ matrix, and we consider the general form

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & -\mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} \quad (176)$$

For simplicity, we assume that the matrix \mathbf{C} is given by $\mathbf{C} = \varepsilon \mathbf{M}'$, where the matrix \mathbf{M}' is attached to the norm $\|\cdot\|_Y$ as in (103). Clearly, the results will apply, almost unchanged, to a symmetric positive definite matrix having maximum and minimum eigenvalue of order ε . We have the following result.

Theorem 3. Let $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ satisfy the system

$$\begin{cases} \mathbf{A}\mathbf{x} + \mathbf{B}^T\mathbf{y} = \mathbf{f} \\ \mathbf{B}\mathbf{x} - \varepsilon\mathbf{M}'\mathbf{y} = \mathbf{g} \end{cases} \quad (177)$$

Assume that \mathbf{A} is symmetric and positive semidefinite, and that the continuity condition (101), the dual norm assumptions (105), the *inf-sup* (112) and the *elker* condition (159) are satisfied. Then, we have

$$\|\mathbf{x}\|_X \leq \frac{\beta^2 + 4\varepsilon M_a}{\alpha_0 \beta^2} \|\mathbf{f}\|_F + \frac{2M_a^{1/2}}{\alpha_0^{1/2} \beta} \|\mathbf{g}\|_G \quad (178)$$

and

$$\|\mathbf{y}\|_Y \leq \frac{2M_a^{1/2}}{\alpha_0^{1/2} \beta} \|\mathbf{f}\|_F + \frac{4M_a}{M_a \varepsilon + \beta^2} \|\mathbf{g}\|_G \quad (179)$$

Proof. The proof can be performed with arguments similar to the ones used in the previous stability proofs, but using more technicalities. For simplicity, we are going to give only a sketch, treating separately the two cases $\mathbf{f} = \mathbf{0}$ and $\mathbf{g} = \mathbf{0}$.

• **The case $\mathbf{f} = \mathbf{0}$.** We set $\tilde{\mathbf{x}} = \mathbf{L}(\mathbf{g} + \varepsilon \mathbf{M}'\mathbf{y})$ and $\mathbf{x}_0 = \mathbf{x} - \tilde{\mathbf{x}}$. Proceeding exactly as in the proof of Theorem 1 (Step 4), we obtain inequality (140):

$$\|\mathbf{y}\|_Y \leq \frac{M_a^{1/2}}{\beta} (\mathbf{x}^T \mathbf{A} \mathbf{x})^{1/2} \quad (180)$$

Then, we multiply the first equation of (176) times \mathbf{x}^T and substitute the value of \mathbf{y} obtained from the second equation. We have

$$\mathbf{x}^T \mathbf{A} \mathbf{x} + \frac{1}{\varepsilon} [(\mathbf{M}')^{-1} (\mathbf{B} \mathbf{x} - \mathbf{g})]^T \mathbf{B} \mathbf{x} = \mathbf{0} \quad (181)$$

Using the fact that $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$, we easily deduce that

$$\|\mathbf{B} \mathbf{x}\|_G \leq \|\mathbf{g}\|_G \quad (182)$$

This implies

$$\|\tilde{\mathbf{x}}\|_X \leq \frac{1}{\beta} \|\mathbf{B} \mathbf{x}\|_G \leq \frac{1}{\beta} \|\mathbf{g}\|_G \quad (183)$$

We now multiply the first equation times \mathbf{x}_0^T , and we have $\mathbf{x}_0^T \mathbf{A} \mathbf{x} = 0$. We can then use (109) to get

$$\mathbf{x}_0^T \mathbf{A} \mathbf{x}_0 = -\mathbf{x}_0^T \mathbf{A} \tilde{\mathbf{x}} \leq (\mathbf{x}_0^T \mathbf{A} \mathbf{x}_0)^{1/2} (\tilde{\mathbf{x}}^T \mathbf{A} \tilde{\mathbf{x}})^{1/2} \quad (184)$$

Simplifying by $(\mathbf{x}_0^T \mathbf{A} \mathbf{x}_0)^{1/2}$ and using (183), we obtain

$$\mathbf{x}_0^T \mathbf{A} \mathbf{x}_0 \leq \tilde{\mathbf{x}}^T \mathbf{A} \tilde{\mathbf{x}} \leq M_a \|\tilde{\mathbf{x}}\|_X^2 \leq \frac{M_a}{\beta^2} \|\mathbf{g}\|_G^2 \quad (185)$$

Using $\mathbf{x} = \mathbf{x}_0 + \tilde{\mathbf{x}}$, and then again (109) and (183), we obtain

$$\mathbf{x}^T \mathbf{A} \mathbf{x} \leq \frac{4M_a}{\beta^2} \|\mathbf{g}\|_G^2 \quad (186)$$

that inserted in (180) gives an estimate for \mathbf{y}

$$\|\mathbf{y}\|_Y \leq \frac{2M_a}{\beta^2} \|\mathbf{g}\|_G \quad (187)$$

On the other hand, using the *elker* condition (159), estimates (183), (185), and (186) we have

$$\begin{aligned} \|\mathbf{x}\|_X &\leq \|\mathbf{x}_0\|_X + \|\tilde{\mathbf{x}}\|_X \leq \left(\frac{M_a^{1/2}}{\alpha_0^{1/2} \beta} + \frac{1}{\beta} \right) \|\mathbf{g}\|_G \\ &= \frac{M_a^{1/2} + \alpha_0^{1/2}}{\alpha_0^{1/2} \beta} \|\mathbf{g}\|_G \leq \frac{2M_a^{1/2}}{\alpha_0^{1/2} \beta} \|\mathbf{g}\|_G \quad (188) \end{aligned}$$

However, we note that using the second equation we might have another possible estimate for \mathbf{y} :

$$\|\mathbf{y}\|_Y \leq \frac{1}{\varepsilon} \|\mathbf{B} \mathbf{x} - \mathbf{g}\|_G \leq \frac{2}{\varepsilon} \|\mathbf{g}\|_G \quad (189)$$

We can combine (187) and (189) into

$$\|\mathbf{y}\|_Y \leq \min \left\{ \frac{2}{\varepsilon}, \frac{2M_a}{\alpha_0^{1/2} \beta} \right\} \|\mathbf{g}\|_G \leq \frac{4M_a}{M_a \varepsilon + \beta^2} \|\mathbf{g}\|_G \quad (190)$$

• **The case $\mathbf{g} = \mathbf{0}$.** We set this time $\tilde{\mathbf{x}} = \mathbf{L}(\varepsilon \mathbf{M}'\mathbf{y})$ and again $\mathbf{x}_0 = \mathbf{x} - \tilde{\mathbf{x}}$. From (150), we have as usual

$$\|\tilde{\mathbf{x}}\|_X \leq \frac{1}{\beta} \|\mathbf{B} \tilde{\mathbf{x}}\|_G = \frac{1}{\beta} \|\mathbf{B} \mathbf{x}\|_G \quad (191)$$

Multiplying the first equation by \mathbf{x}_0^T , we have $\mathbf{x}_0^T \mathbf{A} \mathbf{x} = \mathbf{x}_0^T \mathbf{f}$ that gives, using (159) and (109)

$$\mathbf{x}_0^T \mathbf{A} \mathbf{x}_0 \leq \frac{1}{\alpha_0^{1/2}} \|\mathbf{f}\|_F (\mathbf{x}_0^T \mathbf{A} \mathbf{x}_0)^{1/2} + (\mathbf{x}_0^T \mathbf{A} \mathbf{x}_0)^{1/2} (\tilde{\mathbf{x}}^T \mathbf{A} \tilde{\mathbf{x}})^{1/2} \quad (192)$$

and finally,

$$(\mathbf{x}_0^T \mathbf{A} \mathbf{x}_0)^{1/2} \leq \frac{1}{\alpha_0^{1/2}} \|\mathbf{f}\|_F + (\tilde{\mathbf{x}}^T \mathbf{A} \tilde{\mathbf{x}})^{1/2} \quad (193)$$

In particular, using once more, (109), (193), and (191), we obtain

$$\begin{aligned} |\mathbf{x}_0^T \mathbf{A} \tilde{\mathbf{x}}| &\leq \frac{1}{\alpha_0^{1/2}} \|\mathbf{f}\|_F (\tilde{\mathbf{x}}^T \mathbf{A} \tilde{\mathbf{x}})^{1/2} + \tilde{\mathbf{x}}^T \mathbf{A} \tilde{\mathbf{x}} \\ &\leq \frac{M_a^{1/2}}{\alpha_0^{1/2} \beta} \|\mathbf{f}\|_F \|\mathbf{B} \mathbf{x}\|_G + \tilde{\mathbf{x}}^T \mathbf{A} \tilde{\mathbf{x}} \quad (194) \end{aligned}$$

Take now the product of the first equation times $\tilde{\mathbf{x}}^T$ and using $\mathbf{y} = \varepsilon^{-1} (\mathbf{M}')^{-1} \mathbf{B} \mathbf{x}$ from the second equation, we have $\tilde{\mathbf{x}}^T \mathbf{B}^T \mathbf{y} = \varepsilon^{-1} \tilde{\mathbf{x}}^T \mathbf{B}^T (\mathbf{M}')^{-1} \mathbf{B} \mathbf{x} = \varepsilon^{-1} \|\mathbf{B} \mathbf{x}\|_G^2$. Hence,

$$\tilde{\mathbf{x}}^T \mathbf{A} \mathbf{x} + \frac{1}{\varepsilon} \|\mathbf{B} \mathbf{x}\|_G^2 = \tilde{\mathbf{x}}^T \mathbf{f} \leq \frac{1}{\beta} \|\mathbf{f}\|_F \|\mathbf{B} \mathbf{x}\|_G \quad (195)$$

Using $\tilde{\mathbf{x}}^T \mathbf{A} \mathbf{x} = \tilde{\mathbf{x}}^T \mathbf{A} \tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T \mathbf{A} \mathbf{x}_0$ and the estimate (194) in (195), we deduce

$$\frac{1}{\varepsilon} \|\mathbf{B} \mathbf{x}\|_G^2 \leq \frac{1}{\beta} \|\mathbf{f}\|_F \|\mathbf{B} \mathbf{x}\|_G + \frac{M_a^{1/2}}{\alpha_0^{1/2} \beta} \|\mathbf{f}\|_F \|\mathbf{B} \mathbf{x}\|_G \quad (196)$$

that finally gives

$$\|\mathbf{B} \mathbf{x}\|_G \leq \varepsilon \left(\frac{1}{\beta} + \frac{M_a^{1/2}}{\alpha_0^{1/2} \beta} \right) \|\mathbf{f}\|_F \leq \frac{2\varepsilon M_a^{1/2}}{\alpha_0^{1/2} \beta} \|\mathbf{f}\|_F \quad (197)$$

which is a crucial step in our proof. Indeed, from (197) and the second equation, we obtain our estimate for \mathbf{y}

$$\|\mathbf{y}\|_Y \leq \frac{1}{\varepsilon} \|\mathbf{B} \mathbf{x}\|_G \leq \frac{2M_a^{1/2}}{\alpha_0^{1/2} \beta} \|\mathbf{f}\|_F \quad (198)$$

From (191) and (197), we have

$$\|\tilde{\mathbf{x}}\|_X \leq \frac{1}{\beta} \|\mathbf{B} \mathbf{x}\|_G \leq \frac{2\varepsilon M_a^{1/2}}{\alpha_0^{1/2} \beta^2} \|\mathbf{f}\|_F \quad (199)$$

Finally, from (159), (193), and (199), we obtain

$$\begin{aligned} \|\mathbf{x}_0\|_X &\leq \frac{1}{\alpha_0^{1/2}} (\mathbf{x}_0^T \mathbf{A} \mathbf{x}_0)^{1/2} \leq \left(\frac{1}{\alpha_0} + \frac{2\varepsilon M_a}{\alpha_0 \beta^2} \right) \|\mathbf{f}\|_F \\ &= \frac{\beta^2 + 2\varepsilon M_a}{\alpha_0 \beta^2} \|\mathbf{f}\|_F \quad (200) \end{aligned}$$

which together with (199) gives us the estimate for \mathbf{x}

$$\|\mathbf{x}\|_X \leq \left(\frac{2\varepsilon M_a^{1/2}}{\alpha_0^{1/2} \beta^2} + \frac{2\varepsilon M_a}{\alpha_0 \beta^2} \right) \|\mathbf{f}\|_F \leq \frac{4\varepsilon M_a + \beta^2}{\alpha_0 \beta^2} \|\mathbf{f}\|_F \quad (201)$$

Collecting (190), (188), (198), and (201), we have the result. \square

Remark 6. We notice that the dependence of the stability constants upon α_0 and β in Theorem 3 are optimal, as shown by the system

$$\begin{bmatrix} 2a & \sqrt{a} & -\sqrt{a} & 0 & 0 \\ \sqrt{a} & 2 & 1 & b & 0 \\ -\sqrt{a} & 1 & 2 & 0 & b \\ 0 & b & 0 & -\varepsilon & 0 \\ 0 & 0 & b & 0 & -\varepsilon \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2f \\ 0 \\ 0 \\ 0 \\ 2g \end{bmatrix} \quad (202)$$

Indeed, we have $\alpha_0 = 2a$, $\beta = b$, and the solution is given by

$$x_1 = \frac{f(b^2 + \varepsilon)}{ab^2} + \frac{g}{a^{1/2}b}, \quad x_2 = -\frac{fe}{a^{1/2}b^2} - \frac{3ge}{b(3\varepsilon + b^2)},$$

$$x_3 = \frac{f\varepsilon}{a^{1/2}b^2} + \frac{g(3\varepsilon + 2b^2)}{b(3\varepsilon + b^2)},$$

$$y_1 = -\frac{f}{a^{1/2}b} - \frac{3g}{3\varepsilon + b^2}, \quad y_2 = \frac{f}{a^{1/2}b} - \frac{3g}{3\varepsilon + b^2}$$

Remark 7. It is also worth noticing that assuming full ellipticity of the matrix \mathbf{A} as in (125) (instead of ellipticity only in the kernel as we did here) would improve the estimates for \mathbf{x} . In particular, we could obtain estimates that do not degenerate when β goes to zero, as far as ε remains strictly positive. For the case $f = 0$, this is immediate from the estimate of \mathbf{y} (190): from the first equation, we have easily

$$\|\mathbf{x}\| \leq \frac{1}{\alpha} M_b \|\mathbf{y}\|_F \leq \frac{4M_a M_b}{\alpha(M_a \varepsilon + \beta^2)} \|\mathbf{g}\|_G \quad (203)$$

In the case $\mathbf{g} = 0$, we can combine the two equations to get

$$\mathbf{x}^T \mathbf{A} \mathbf{x} + \varepsilon \|\mathbf{y}\|_F^2 = \mathbf{x}^T \mathbf{f} \quad (204)$$

that gives (always using (125))

$$\|\mathbf{x}\|_X \leq \frac{1}{\alpha} \|\mathbf{f}\|_F \quad (205)$$

that then gives

$$\|\mathbf{y}\|_F \leq \frac{1}{\varepsilon} \|\mathbf{B} \mathbf{x}\|_G \leq \frac{M_b}{\varepsilon \alpha} \|\mathbf{f}\|_F \quad (206)$$

This could be combined with (198) into

$$\|\mathbf{y}\|_F \leq \min \left\{ \frac{M_b}{\varepsilon \alpha}, \frac{2M_a^{1/2}}{\alpha^{1/2} \beta} \right\} \|\mathbf{f}\|_F$$

$$\leq \frac{4M_a^{1/2} M_b}{2M_a^{1/2} \alpha \varepsilon + \alpha^{1/2} \beta M_b} \|\mathbf{f}\|_F \quad (207)$$

Collecting the two cases we have

$$\|\mathbf{x}\|_X \leq \frac{1}{\alpha} \|\mathbf{f}\|_F + \frac{4M_a M_b}{\alpha^{1/2} (M_a \varepsilon + \beta^2)} \|\mathbf{g}\|_G \quad (208)$$

and

$$\|\mathbf{y}\|_F \leq \frac{4M_a^{1/2} M_b}{2M_a^{1/2} \alpha \varepsilon + \alpha^{1/2} \beta M_b} \|\mathbf{f}\|_F + \frac{4M_a}{M_a \varepsilon + \beta^2} \|\mathbf{g}\|_G \quad (209)$$

which do not degenerate for β going to zero.

As announced in the title of the section, systems of the type (176) occur, for instance, in the so-called (\mathbf{u}, π) formulation of nearly incompressible elasticity. Sometimes they are also obtained by penalizing systems of the original type (95) in order to obtain a partial cure in cases in which

β is zero or tending to zero with the meshsize (as it could happen, for instance, for a discretization of Stokes problem that does not satisfy the *inf-sup* condition), in the spirit of Remark 7. Indeed, the (\mathbf{u}, π) formulation of nearly incompressible elasticity, in the case of an isotropic and homogeneous body, could be seen, mathematically, as a perturbation of the Stokes system with $\varepsilon = 1/\lambda$, and the elements to be used are essentially the same.

3.8 Composite matrices

In a certain number of applications, one has to deal with formulations of mixed type where more than two fields are involved. These give rise to matrices that are naturally split as 3×3 or 4×4 (or more) block matrices. For the sake of completeness, we show how the previous theory can often apply almost immediately to these more general cases. As an example, we consider matrices of the type

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T & \mathbf{0} \\ \mathbf{B} & \mathbf{0} & \mathbf{C}^T \\ \mathbf{0} & \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{y}_1 \end{Bmatrix} = \begin{Bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{g} \end{Bmatrix} \quad (210)$$

Matrices of the form (210) are found (among several other applications) in the discretization of formulations of Hu–Washizu type. However, in particular, for elasticity problems, there are no good examples of finite element discretizations of the Hu–Washizu principle that satisfy the following two requirements at the same time: not reducing more or less immediately (in the linear case) to known discretizations of the minimum potential energy or of the Hellinger–Reissner principle, and having been proved to be stable and optimally convergent in a sound mathematical way. Actually, the only way, so far, has been using *stabilized formulations* (see for instance Behr, Franca and Tezduyar, 1993) that we decided to avoid here. Still, we hope that the following brief discussion could also be useful for the possible development of good Hu–Washizu elements in the future.

Coming back to the analysis of (210), we already observed that systems of this type can be reconduced to the general form (95)

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{x} \\ \mathbf{y} \end{Bmatrix} = \begin{Bmatrix} \mathbf{f} \\ \mathbf{g} \end{Bmatrix} \quad (211)$$

after making the following simple identifications:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{C} \end{bmatrix}$$

$$\mathbf{x} = \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{Bmatrix}, \quad \mathbf{y} = \mathbf{y}_1 \quad (212)$$

The stability of system (211) can then be studied using the previous analysis. Sometimes it is, however, more convenient to reach the compact form (211) with a different identification:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B} & \mathbf{C}^T \end{bmatrix}$$

$$\mathbf{x} = \begin{Bmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{Bmatrix}, \quad \mathbf{y} = \mathbf{x}_2 \quad (213)$$

Indeed, in this case, the matrix \mathbf{A} is much simpler. In particular, as it happens quite often in practice, when the original matrix \mathbf{A} in (210) is symmetric and positive semidefinite, the same properties will be shared by \mathbf{A} . We are not going to repeat the theory of the above sections for the extended systems (210). We will just point out the meaning of conditions *elker* and *inf-sup*, applied to the system (212) to (213), in terms of the original matrices \mathbf{A} , \mathbf{B} , and \mathbf{C} .

The kernel of \mathbf{B} , as given in (213), is made of the pairs $(\mathbf{x}_1, \mathbf{y}_1)$ such that

$$\mathbf{B} \mathbf{x}_1 + \mathbf{C}^T \mathbf{y}_1 = \mathbf{0} \quad (214)$$

These include, in particular, all the pairs $(0, \mathbf{y}_1)$, where \mathbf{y}_1 is in the kernel of \mathbf{C}^T :

$$\text{Ker}(\mathbf{C}^T) := \{\mathbf{y}_1 \mid \text{such that } \mathbf{C}^T \mathbf{y}_1 = \mathbf{0}\} \quad (215)$$

There is no hope that the matrix \mathbf{A} , as defined in (213), can be elliptic on those pairs. Hence, we must require that those pairs are actually reduced to the pair $(0, 0)$, that is, we must require that

$$\text{Ker}(\mathbf{C}^T) = \{0\} \quad (216)$$

This does not settle the matter of *elker*, since there are many other pairs $(\mathbf{x}_1, \mathbf{y}_1)$ satisfying (214). As \mathbf{A} acts only on the \mathbf{x}_1 variables, we must characterize the vectors \mathbf{x}_1 such that $(\mathbf{x}_1, \mathbf{y}_1)$ satisfies (214) for some \mathbf{y}_1 . These are

$$\mathcal{K} := \{\mathbf{x}_1 \mid \text{such that } \mathbf{z}^T \mathbf{B} \mathbf{x}_1 = 0 \quad \forall \mathbf{z} \in \text{Ker}(\mathbf{C})\} \quad (217)$$

Hence we have the following result: condition *elker* will hold, for the system (212) to (213) if and only if

$$\exists \tilde{\alpha} > 0 \quad \text{such that } \tilde{\alpha} \|\mathbf{x}_1\|^2 \leq \mathbf{x}_1^T \mathbf{A} \mathbf{x}_1 \quad \forall \mathbf{x}_1 \in \mathcal{K} \quad (218)$$

On the other hand, it is not difficult to see that condition *inf-sup* for (212) to (213) reads

$$\exists \tilde{\beta} > 0 \quad \text{such that } \sup_{(\mathbf{x}_1, \mathbf{y}_1)} \frac{\mathbf{x}_2^T \mathbf{B} \mathbf{x}_1 + \mathbf{x}_2^T \mathbf{C}^T \mathbf{y}_1}{\|\mathbf{x}_1\| + \|\mathbf{y}_1\|} \geq \tilde{\beta} \|\mathbf{x}_2\| \quad \forall \mathbf{x}_2 \quad (219)$$

It is clear that a *sufficient* condition would be to have the *inf-sup* condition to hold for at least one of the two matrices \mathbf{B} , \mathbf{C}^T . In many applications, however, this is too strong a requirement. A weaker condition (although stronger than (219)) can be written as

$$\exists \tilde{\beta} > 0 \quad \text{such that } \tilde{\beta} \|\mathbf{x}_2\| \leq \|\mathbf{C} \mathbf{x}_2\| + \|\mathbf{B}^T \mathbf{x}_2\| \quad \forall \mathbf{x}_2 \quad (220)$$

More generally, many variations are possible, according to the actual structure of the matrices at play.

4 APPLICATIONS

In this section, we give several examples of efficient mixed finite element methods, focusing our attention mostly on the thermal problem (Section 4.1) and on the Stokes equation (Section 4.2). For simplicity, we mainly consider triangular elements, while we briefly discuss their possible extensions to quadrilateral geometries and to three-dimensional cases. Regarding Stokes equation, we point out (as already mentioned) that the same discretization spaces can be profitably used to treat the nearly incompressible elasticity problem, within the context of the (\mathbf{u}, π) formulation (80). We also address a brief discussion on elements for the elasticity problem in the framework of the Hellinger–Reissner principle (Section 4.3).

We finally remark that, for all the schemes that we are going to present, a rigorous stability and convergence analysis has been established, even though we will not detail the proofs.

4.1 Thermal diffusion

We consider the thermal diffusion problem described in Section 2.1 in the framework of the Hellinger–Reissner variational principle. We recall that the discretization of such a problem leads to solve the following algebraic system:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \hat{\mathbf{q}} \\ \hat{\boldsymbol{\theta}} \end{Bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{g} \end{Bmatrix} \quad (221)$$

where

$$\begin{cases} \mathbf{A}_{ij} = \int_{\Omega} [\mathbf{N}_i^T \cdot \mathbf{D}^{-1} \mathbf{N}_j^T] d\Omega, & \hat{\mathbf{q}}_i = \hat{q}_i \\ \mathbf{B}_{i,j} = - \int_{\Omega} [N_i^T \text{div}(\mathbf{N}_j^T)] d\Omega, & \hat{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\theta}}_i \\ \mathbf{g}_i = \int_{\Omega} [N_i^T b] d\Omega \end{cases} \quad (222)$$

Above, N_i^q and N_j^θ are the interpolation functions for the flux q and the temperature θ respectively. Moreover, \hat{q} and $\hat{\theta}$ are the vectors of flux and temperature unknowns, while $i, j = 1, \dots, n$ and $r = 1, \dots, m$, where n and m obviously depend on the chosen approximation spaces as well as on the mesh.

Following the notation of the previous section, the norms for which the *inf-sup* and the *elker* conditions should be checked are (cf. (174) and (175))

$$\|\hat{q}\|_X^2 := \int_{\Omega} \left[(N_i^q \hat{q}_i) \cdot D^{-1} (N_j^q \hat{q}_j) \right] d\Omega + \frac{\ell^2}{k_*} \int_{\Omega} |\operatorname{div} (N_i^q \hat{q}_i)|^2 d\Omega \quad (223)$$

and

$$\|\hat{\theta}\|_Y^2 := \int_{\Omega} |N_r^\theta \hat{\theta}_r|^2 d\Omega \quad (224)$$

where ℓ is some characteristic length of the domain Ω and k_* is some characteristic value of the thermal conductivity.

Before proceeding, we remark the following:

- Since no derivative operator acts on the interpolating functions N_i^q in the matrix B , we are allowed to approximate the temperature θ without requiring any continuity across the elements. On the contrary, the presence of the divergence operator acting on the interpolating functions N_i^q in the matrix B suggests that the normal component of the approximated flux should not exhibit jumps between adjacent elements.
- The full ellipticity for A (i.e. property (125)) typically holds only with a constant $\alpha \approx h^2$, once the norm (223) has been chosen. However, if a method is designed in such a way that

$$\hat{q}_0 = (\hat{q}_i^0)_{i=1}^n \in \operatorname{Ker}(B) \text{ implies } \operatorname{div} (N_i^q \hat{q}_i^0) = 0 \quad (225)$$

the weaker *elker* condition (159) obviously holds with $\alpha_0 = 1$. Condition (225) is verified if, for instance, we insist that

$$\operatorname{Span}\{\operatorname{div} N_i^q; i = 1, \dots, n\} \subseteq \operatorname{Span}\{N_r^\theta; r = 1, \dots, m\} \quad (226)$$

that is, the divergences of all the approximated fluxes are contained in the space of the approximated temperatures. Indeed, condition (226) implies that, for every $\hat{q}_0 \in \operatorname{Ker}(B)$, there exists $\hat{\theta}_0 = (\hat{\theta}_r^0)_{r=1}^m$ such that

$\operatorname{div} (N_i^q \hat{q}_i^0) = -N_r^\theta \hat{\theta}_r^0$. It follows that

$$0 = \hat{\theta}_0^T B \hat{q}_0 = - \int_{\Omega} (N_r^\theta \hat{\theta}_r^0) \operatorname{div} (N_i^q \hat{q}_i^0) d\Omega = \int_{\Omega} |\operatorname{div} (N_i^q \hat{q}_i^0)|^2 d\Omega \quad (227)$$

so that $\operatorname{div} (N_i^q \hat{q}_i^0) = 0$.

Condition (226) can be always achieved by 'enriching' the temperature approximation, if necessary. However, we remark that a careless enlargement of the approximated temperatures can compromise the fulfillment of the *inf-sup* condition (112), as shown in the following easy result.

Proposition 3. Suppose that a given method satisfies condition (226). Then the *inf-sup* condition (112) implies

$$\operatorname{Span}\{\operatorname{div} N_i^q; i = 1, \dots, n\} = \operatorname{Span}\{N_r^\theta; r = 1, \dots, m\} \quad (228)$$

that is, the divergences of all the approximated fluxes coincide with the space of the approximated temperatures.

Proof. By contradiction, suppose that $\operatorname{Span}\{\operatorname{div} N_i^q; i = 1, \dots, n\}$ is strictly contained in $\operatorname{Span}\{N_r^\theta; r = 1, \dots, m\}$. It follows that there exists $\hat{\theta}_\perp = (\hat{\theta}_r^\perp)_{r=1}^m \in \mathbb{R}^m \setminus \{0\}$ such that

$$\hat{q}_\perp^T B^T \hat{\theta}_\perp = - \int_{\Omega} (N_r^\theta \hat{\theta}_r^\perp) \operatorname{div} (N_i^q \hat{q}_i^*) d\Omega = 0 \quad \forall \hat{q}_\perp \in \mathbb{R}^n \quad (229)$$

Therefore,

$$\sup_{\hat{q}_\perp \in \mathbb{R}^n \setminus \{0\}} \frac{\hat{q}_\perp^T B^T \hat{\theta}_\perp}{\|\hat{q}_\perp\|_X} = 0 \quad (230)$$

and the *inf-sup* condition does not hold (cf. (119)). \square

We also remark that the converse of Proposition 3 does not hold, that is, condition (228) is not sufficient for the fulfillment of *inf-sup* (although it does imply *elker*).

From the considerations above, it should be clear that

- degrees of freedom associated with the **normal component** of the approximated flux are needed to guarantee its continuity across adjacent elements;
- the satisfaction of both the *elker* and the *inf-sup* condition requires a careful and well-balanced choice of the interpolating fields.

In the following, we are going to present several elements designed accordingly to the guidelines above, all satisfying property (228).

4.1.1 Triangular elements

Throughout this section, we will always suppose that the domain $\Omega \subset \mathbb{R}^2$, on which the thermal problem is posed, is decomposed by means of a triangular mesh \mathcal{T}_h with meshsize h . Moreover, we define \mathcal{E}_h as the set of all the edges of the triangles in \mathcal{T}_h .

• **The $RT_0 - P_0$ element.** We now introduce the simplest triangular element proposed for thermal problems. For the discretization of the thermal flux q , we take the so-called *lowest-order Raviart-Thomas element* (RT_0 element), presented in Raviart and Thomas (1977); accordingly, the approximated flux q^h is described as a *piecewise linear* (vectorial) field such that

- the normal component $q^h \cdot n$ is constant on each edge e of \mathcal{E}_h ;
- the normal component $q^h \cdot n$ is continuous across each edge e of \mathcal{E}_h .

To approximate the temperature, we simply use *piecewise constant functions* in each element (P_0 element).

On the generic triangle $T \in \mathcal{T}_h$, a set of element degrees of freedom for q^h is given by its 3 normal fluxes on the edges of the triangle, that is,

$$\int_e q^h \cdot n \, ds \quad \forall e \text{ edge of } T \quad (231)$$

Therefore, the space for the element approximation of q has dimension 3 and a basis is obtained by considering the (vectorial) shape functions

$$N_k^q = N_k^q(x, y) = \frac{1}{2 \operatorname{Area}(T)} \begin{cases} x - x_k \\ y - y_k \end{cases} \quad k = 1, 2, 3 \quad (232)$$

Above, $(x_k, y_k)^T$ denotes the position vector of the k th vertex (local numbering) of the triangle T .

We also remark that, because of (232), q^h can be locally described by

$$q^h = p_0 + p_0 \begin{Bmatrix} x \\ y \end{Bmatrix} = \begin{Bmatrix} a_0 + p_0 x \\ b_0 + p_0 y \end{Bmatrix} \quad (233)$$

where $a_0, b_0, p_0 \in \mathbb{R}$.

As far as the approximated temperature is concerned, an element basis for θ^h is given by the shape function

$$N^0 = N^0(x, y) = 1 \quad (234)$$

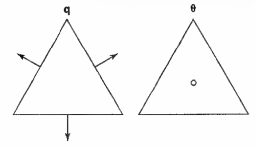


Figure 1. Degrees of freedom for $RT_0 - P_0$ element.

The element degrees of freedom for both q^h and θ^h are schematically depicted in Figure 1.

• **The $RT_k - P_k$ family.** We now present the extension to higher orders of the $RT_0 - P_0$ method just described (cf. Raviart and Thomas, 1977). Given an integer $k \geq 1$ and using the definition introduced in Nedelec (1980), for the flux q^h , we take a field such that (RT_k element) on each triangle T of \mathcal{T}_h , we have

$$q^h = p_k(x, y) + p_k(x, y) \begin{Bmatrix} x \\ y \end{Bmatrix} \quad (235)$$

where $p_k(x, y)$ (resp. $p_k(x, y)$) is a vectorial (resp. scalar) polynomial of degree at most k . Moreover, we require that the **normal component** $q^h \cdot n$ is *continuous* across each edge e of \mathcal{E}_h . This can be achieved by selecting the following element degrees of freedom:

- the moments of order up to k of $q^h \cdot n$ on the edges of T ;
- the moments of order up to $k - 1$ of q^h on T .

For the discretized temperature θ^h , we take *piecewise polynomials of degree at most k* (P_k element).

The element degrees of freedom for the choice $k = 1$ are shown in Figure 2.

• **The $BDM_1 - P_0$ element.** Another method, widely used to treat the thermal diffusion problem, arises from the approximation of the flux q by means of the so-called

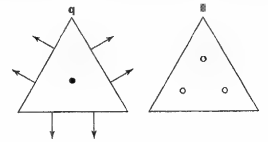


Figure 2. Degrees of freedom for $RT_1 - P_1$ element.

lowest-order Brezzi–Douglas–Marini element (BDM_1 element), proposed and analyzed in Brezzi *et al.* (1985). It consists in discretizing \mathbf{q} by vectorial functions \mathbf{q}^h such that

- \mathbf{q}^h is linear in triangle T of \mathcal{T}_h ;
- the normal component $\mathbf{q}^h \cdot \mathbf{n}$ is continuous across each edge e of \mathcal{E}_h .

For the approximated temperature θ^h , we again use piecewise constant functions on each triangle.

Focusing on the generic triangle $T \in \mathcal{T}_h$, we remark that the approximation space for \mathbf{q} has dimension 6, since full linear polynomials are employed. A suitable set of element degrees of freedom is provided by the moments up to order 1 of the normal fluxes $\mathbf{q}^h \cdot \mathbf{n}$ across each edge e of T , explicitly given by the values

$$\left\{ \begin{array}{l} \int_e \mathbf{q}^h \cdot \mathbf{n} \, ds \\ \int_s s \mathbf{q}^h \cdot \mathbf{n} \, ds \end{array} \right. \quad (236)$$

where s is a local coordinate on e ranging from -1 to 1 .

The element degrees of freedom for the resulting method are shown in Figure 3.

• *The $BDM_{k+1} - P_k$ family.* As for the $RT_0 - P_0$ scheme, also the $BDM_{k+1} - P_0$ finite element method is the lowest order representative of a whole class. Indeed, given an integer $k \geq 1$, we can select the approximations presented in Brezzi *et al.* (1985).

For the discretized flux \mathbf{q}^h , the normal component $\mathbf{q}^h \cdot \mathbf{n}$ is continuous across each edge e of \mathcal{E}_h . Moreover, \mathbf{q}^h is a vectorial polynomial of degree at most $k+1$ on each triangle T of \mathcal{T}_h (BDM_{k+1} element). Also, in this case, the continuity of the normal component can be obtained by a proper choice of the degrees of freedom.

For the approximated temperature θ^h , we use the discontinuous P_k element. Figure 4 shows the element degrees of freedom for the case $k=1$.

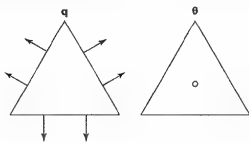


Figure 3. Degrees of freedom for $BDM_1 - P_0$ element.

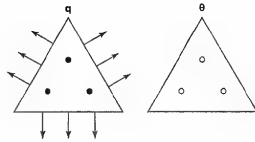


Figure 4. Degrees of freedom for $BDM_2 - P_1$ element.

4.1.2 Quadrilateral elements

We now briefly consider the extension of some of the methods presented in the previous section to quadrilateral meshes. In this case, we define our approximating spaces on the reference element $\tilde{K} = [-1, 1]^2$ equipped with local coordinates (ξ, η) . As far as the flux is concerned, the corresponding approximation space on each physical element K must be obtained through the use of a suitable transformation that preserves the normal component of vectorial functions. This is accomplished by the following (contravariant) Piola's transformation of vector fields. Suppose that

$$\mathbf{F}: \tilde{K} \longrightarrow K; \quad (x, y) = \mathbf{F}(\xi, \eta)$$

is an invertible map from \tilde{K} onto K , with Jacobian matrix $\mathbf{J}(\xi, \eta)$. Given a vector field $\mathbf{q} = \mathbf{q}(\xi, \eta)$ on \tilde{K} , its Piola's transform $\mathcal{P}(\mathbf{q}) = \mathcal{P}(\mathbf{q})(x, y)$ is the vector field on K , defined by

$$\mathcal{P}(\mathbf{q})(x, y) := \frac{1}{J(\xi, \eta)} \mathbf{J}(\xi, \eta) \mathbf{q}(\xi, \eta); \quad (x, y) = \mathbf{F}(\xi, \eta)$$

where $J(\xi, \eta) = |\det \mathbf{J}(\xi, \eta)|$. Therefore, if

$$\mathbf{Q}(\tilde{K}) = \text{Span}(\mathbf{q}_i^*); \quad i = 1, \dots, n_d$$

is an n_d -dimensional flux approximation space defined on the reference element \tilde{K} , the corresponding space on the physical element K will be

$$\mathbf{Q}(K) = \text{Span}(\mathcal{P}(\mathbf{q}_i^*)); \quad i = 1, \dots, n_d$$

• *The $RT_0 - P_0$ element.* In the reference element \tilde{K} , we prescribe the approximated flux \mathbf{q}^h as (RT_0 element)

$$\mathbf{q}^h = \begin{Bmatrix} a + b\xi \\ c - d\eta \end{Bmatrix}, \quad a, b, c, d \in \mathbb{R} \quad (237)$$

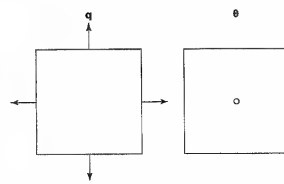


Figure 5. Degrees of freedom for $RT_0 - P_0$ element.

Because of (237), it is easily seen that the four values

$$\int_e \mathbf{q}^h \cdot \mathbf{n} \, ds \quad \forall e \text{ edge of } \tilde{K} \quad (238)$$

can be chosen as a set of degrees of freedom. Moreover, $\text{div } \mathbf{q}^h$ is constant in \tilde{K} , suggesting the choice of a constant approximated temperature θ^h in \tilde{K} (P_0 element). The degrees of freedom for both \mathbf{q}^h and θ^h are shown in Figure 5.

• *The $BDM_{1[1]} - P_0$ element.* For the discrete flux \mathbf{q}^h on \tilde{K} , we take a field such that ($BDM_{1[1]}$ element)

$$\begin{aligned} \mathbf{q}^h &= \mathbf{p}_1(\xi, \eta) + a \begin{Bmatrix} \xi^2 \\ -2\xi\eta \end{Bmatrix} + b \begin{Bmatrix} 2\xi\eta \\ -\eta^2 \end{Bmatrix} \\ &= \mathbf{p}_1(\xi, \eta) + a(\nabla(\xi^2\eta))^{\perp} + b(\nabla(\xi\eta^2))^{\perp} \end{aligned} \quad (239)$$

Above, $\mathbf{p}_1(\xi, \eta)$ is a vectorial linear polynomial, and a, b are real numbers. This space is carefully designed in order to have

- $\mathbf{q}^h \cdot \mathbf{n}$ linear on each edge e of \tilde{K} .
- $\text{div } \mathbf{q}^h$ constant in \tilde{K} .

Again, for the approximated temperature θ^h , we take constant functions (P_0 element). The element degrees of freedom for both \mathbf{q}^h and θ^h are shown in Figure 6.

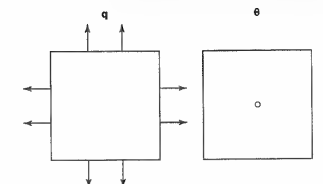


Figure 6. Degrees of freedom for $BDM_{1[1]} - P_0$ element.

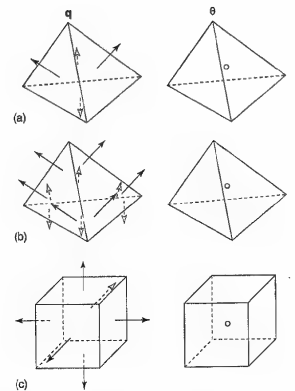


Figure 7. 3-D elements for the thermal problem.

(see Nedelec, 1980):

$$\mathbf{q}^h|_T = \mathbf{p}_0 + p_0 \begin{Bmatrix} x \\ y \\ z \end{Bmatrix} = \begin{Bmatrix} a_0 + p_0 x \\ b_0 + p_0 y \\ c_0 + p_0 z \end{Bmatrix} \quad (240)$$

Therefore, in each tetrahedron T , the space for the approximated flux has dimension 4 and the degrees of freedom are precisely the values $\int_f \mathbf{q}^h \cdot \mathbf{n} \, ds$ on each tetrahedron face f .

The three-dimensional version of the $BDM_1 - P_0$ element (cf. Figure 3) is shown in Figure 7(b). The approximated temperature is still piecewise constant, while the discretized flux $\mathbf{q}^h|_T$ is a fully linear vectorial function.

We also present the extension of the $RT_{01} - P_0$ to the case of cubic geometry, as depicted in Figure 7(c).

Remark 8. We conclude our discussion on the thermal problem by noticing the obvious fact that the linear system (221) has an indefinite matrix, independent of the chosen approximation spaces. This is a serious source of trouble. For the discretizations considered above, we can however overcome this drawback. Following Fraeijns de Veubeke (1965), one can first work with fluxes that are *totally discontinuous*, forcing back the continuity of the normal components by means of suitable *interelement Lagrange multipliers*, whose physical meaning comes out to be 'generalized temperatures' (for instance, approximations of the temperature mean value on each edge). As the fluxes are now discontinuous, it is possible to eliminate them by static condensation at the element level. This will give a system involving only the temperatures and the interelement multipliers. At this point, however, it becomes possible to eliminate the temperatures as well (always at the element level), leaving a final system that involves only the multipliers. This final system has a symmetric and positive definite matrix, a very useful property from the computational point of view. For a detailed discussion about these ideas, we refer to Arnold and Brezzi (1985), Marini (1985), and Brezzi *et al.* (1986, 1987, 1988). For another way to eliminate the flux variables (although with some geometrical restrictions) see also Baranger, Maitre and Oudin (1996). For yet another procedure to reduce the number of unknowns in (221) and getting a symmetric positive definite matrix, see Alotto and Perugia (1999).

4.2 Stokes equation

As detailed in Section 2.2, the discretization of the Stokes problem leads to solving the following algebraic system:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} \quad (241)$$

where

$$\begin{cases} \mathbf{A}_{ij} = \mu \int_{\Omega} [\nabla \mathbf{N}_i^u : \nabla \mathbf{N}_j^u] d\Omega, & \hat{\mathbf{u}}_i = \hat{\mathbf{u}}_i \\ \mathbf{B}_{ij} = - \int_{\Omega} [N_j^p \operatorname{div}(\mathbf{N}_i^u)] d\Omega, & \hat{\mathbf{p}}_i = \hat{\mathbf{p}}_i \\ \mathbf{f}_i = \int_{\Omega} [\mathbf{N}_i^u \cdot \mathbf{b}] d\Omega \end{cases} \quad (242)$$

Above, \mathbf{N}_i^u and N_j^p are the interpolation functions for the velocity \mathbf{u} and the pressure p respectively. Also, $\hat{\mathbf{u}}$ and $\hat{\mathbf{p}}$ are the vectors containing the velocity and the pressure unknowns. In the sequel, we will always consider the case of homogeneous boundary conditions for the velocity field along the whole boundary $\partial\Omega$. As a consequence, the pressure field is determined only up to a constant. Uniqueness can, however, be recovered, for instance, by insisting that the pressure has zero mean value over the domain Ω or by fixing its value at a given point.

We also remark that, since there is no derivative of N_j^p in the definition of the matrix \mathbf{B} , both continuous and discontinuous pressure approximations can be chosen. On the contrary, the symmetric gradients of \mathbf{N}_i^u entering in the matrix \mathbf{A} suggest that the approximated velocities should be continuous across adjacent elements.

If we introduce the norms

$$\|\hat{\mathbf{u}}\|_X^2 := \mu \int_{\Omega} |\nabla(\mathbf{N}_i^u \hat{\mathbf{u}}_i)|^2 d\Omega \quad (243)$$

and

$$\|\hat{\mathbf{p}}\|_Y^2 := \int_{\Omega} |N_j^p \hat{\mathbf{p}}_j|^2 d\Omega \quad (244)$$

the continuity conditions (101) and the ellipticity condition (125) of the previous section are clearly satisfied, namely, with $M_u = 1$, $M_p = \sqrt{d/\mu}$, and $\alpha = 1$. Therefore, a stable method is achieved provided the only *inf-sup* condition (112) is fulfilled.

4.2.1 Triangular elements with continuous pressure interpolation

In this section, we describe some stable triangular element for which the pressure field is interpolated by means of continuous functions.

• **The MINI element.** Given a triangular mesh \mathcal{T}_h of Ω , for the approximated velocity \mathbf{u}^h we require that (cf. Arnold, Brezzi and Fortin, 1984)

- for each $T \in \mathcal{T}_h$, the two components of \mathbf{u}^h are the sum of a linear function plus a standard cubic bubble function;
- the two components of \mathbf{u}^h are globally continuous functions on Ω .

Concerning the discretized pressure p^h , we simply take piecewise linear and globally continuous functions.

For the generic element $T \in \mathcal{T}_h$, the elemental degrees of freedom for \mathbf{u}^h are its (vectorial) values at the triangle vertices and barycenter. A basis for the element approximation space of each component of \mathbf{u}^h can be obtained by

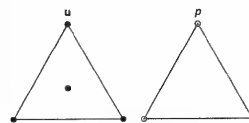


Figure 8. Degrees of freedom for MINI element.

considering the following four shape functions:

$$\begin{cases} N_k = N_k(x, y) = \lambda_k & k = 1, 2, 3 \\ N_b = N_b(x, y) = 27\lambda_1\lambda_2\lambda_3 \end{cases} \quad (245)$$

where $\{\lambda_k = \lambda_k(x, y), k = 1, 2, 3\}$ denote the usual area coordinates on T .

Furthermore, a set of elemental degrees of freedom for p^h is given by its values at the triangle vertices, while the three shape functions to be used are obviously

$$N_k = \lambda_k \quad k = 1, 2, 3 \quad (246)$$

The element degrees of freedom for both \mathbf{u}^h and p^h are schematically depicted in Figure 8. We finally remark that the bubble functions for the velocity are internal modes, so that they can be eliminated on the element level by means of the so-called *static condensation* procedure (cf. Hughes, 1987, for instance). As a consequence, these additional degrees of freedom do not significantly increase the computational costs.

• **The Hood-Taylor elements.** These elements arise from the experimental evidence that using a velocity approximation of one degree higher than the approximation for pressure gave reliable results (cf. Hood and Taylor, 1973). We are therefore led to consider, for each integer k with $k \geq 1$, the following interpolation fields.

The approximated velocity \mathbf{u}^h is such that

- for each $T \in \mathcal{T}_h$, the two components of \mathbf{u}^h are polynomials of degree at most $k+1$;
- the two components of \mathbf{u}^h are globally continuous functions on Ω .

For the approximated pressure p^h , we ask that

- for each $T \in \mathcal{T}_h$, p^h is a polynomial of degree at most k ;
- p^h is a globally continuous function on Ω .

Figure 9 shows the \mathbf{u}^h and p^h element degrees of freedom, for the lowest-order Hood-Taylor method (i.e. $k = 1$).

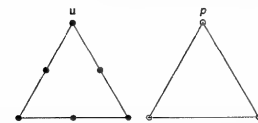


Figure 9. Degrees of freedom for the lowest-order Hood-Taylor element.

Remark 9. A first theoretical analysis of the lowest-order Hood-Taylor method ($k = 1$) was developed in Bercovier and Pironneau (1977), later improved in Verfürth (1984). The case $k = 2$ was treated in Brezzi and Falk (1991), while an analysis covering every choice of k was presented in Boffi (1994). We also remark that the *discontinuous pressure* version of the Hood-Taylor element typically results in an unstable method. However, stability can be recovered by imposing certain restrictions on the mesh for $k \geq 3$ (see Vogelius, 1983; Scott and Vogelius, 1985), or by taking advantage of suitable stabilization procedures for $k \geq 1$ (see Mansfield, 1982; Boffi, 1995).

• **The $(P_1\text{-iso-}P_2) - P_1^c$ element.** This is a 'composite' element whose main advantage is the shape function simplicity. We start by considering a triangular mesh \mathcal{T}_h with meshsize h . From \mathcal{T}_h , we build another finer mesh $\mathcal{T}_{h/2}$ by splitting each triangle T of \mathcal{T}_h into four triangles using the edge midpoints of T , as sketched in Figure 10.

The approximated velocity \mathbf{u}^h is now defined using the finer mesh $\mathcal{T}_{h/2}$ according to the following prescriptions:

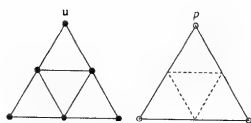
- for each triangle of $\mathcal{T}_{h/2}$, the two components of \mathbf{u}^h are linear functions;
- the two components of \mathbf{u}^h are globally continuous functions on Ω .

On the other hand, the interpolated pressure p^h is piecewise linear in the coarser mesh \mathcal{T}_h , and globally continuous on Ω .

For every triangle T' of finer mesh $\mathcal{T}_{h/2}$, the degrees of freedom of \mathbf{u}^h are its values at the vertices, while an element basis is given by taking the shape functions $N_k = \lambda_k$ ($k = 1, 2, 3$) relative to T' .



Figure 10. Splitting of a triangle $T \in \mathcal{T}_h$.

Figure 11. Degrees of freedom for $(P_1\text{-iso-}P_2) - P_1^c$ element.

Instead, by considering the generic triangle T of the coarser mesh \mathcal{T}_h , the point values at the three vertices provide a set of degrees of freedom for p^h . Therefore, the shape functions $N_k = \lambda_k$ ($k = 1, 2, 3$), relative to T , can be chosen as a basis for the element pressure approximation.

The element degrees of freedom for both u^h and p^h are schematically depicted in Figure 11.

Remark 10. A popular way to solve system (241) consists in using a penalty method. More precisely, instead of (241), one considers the perturbed system

$$\begin{bmatrix} A & B^T \\ B & -\sigma C \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix} \quad (247)$$

where the 'mass' matrix C is defined by

$$C|_T = \int_T [N_i^T N_j^T] d\Omega \quad (248)$$

and $\sigma > 0$ is a 'small' parameter. In the case of discontinuous pressure approximations, the pressure unknowns can be eliminated from (247) on the element level, leading therefore to the following system for \hat{u} :

$$(A + \sigma^{-1} B^T C^{-1} B) \hat{u} = f \quad (249)$$

with C 'easy-to-invert' (namely, block diagonal). When continuous pressure approximations are considered, the inverse of C is in general a full matrix, so that the elimination of the pressure unknowns seems impossible on the element level. We have, however, the right to choose a different penalizing term in (247): for instance, we could replace C by a diagonal matrix \tilde{C} , obtained from C by a suitable mass lumping procedure (cf. Hughes, 1987). The pressure elimination becomes now easy to perform, leading to (cf. (249))

$$(A + \sigma^{-1} B^T \tilde{C}^{-1} B) \hat{u} = f \quad (250)$$

A drawback of this approach, however, not so serious for low-order schemes, stands in a larger bandwidth for the matrix $(A + \sigma^{-1} B^T \tilde{C}^{-1} B)$. For more details about this strategy, we refer to Arnold, Brezzi and Fortin (1984).

4.2.2 Triangular elements with discontinuous pressure interpolation

In this section, we describe some stable triangular element for which the pressure field is interpolated by means of discontinuous functions. It is worth noticing that all these elements have velocity degrees of freedom associated with the element edges. This feature is indeed of great help in proving the *inf-sup* condition for elements with discontinuous pressure interpolation (cf. Remark 16).

• **The Crouzeix-Raviart element.** Our first example of discontinuous pressure elements is the one proposed and analyzed in Crouzeix and Raviart (1973). It consists in choosing the approximated velocity u^h such that

- for each $T \in \mathcal{T}_h$, the two components of u^h are the sum of a quadratic function plus a standard cubic bubble function;
- the two components of u^h are globally continuous functions on Ω .

Moreover, for the discretized pressure p^h , we simply take the piecewise linear functions, without requiring any continuity between adjacent elements.

The elemental approximation of each component of u^h can be described by means of the following seven shape functions

$$\begin{cases} N_k = \lambda_k & k = 1, 2, 3 \\ N_4 = 4\lambda_1\lambda_2, & N_5 = 4\lambda_1\lambda_3, & N_6 = 4\lambda_1\lambda_2 \\ N_7 = 27\lambda_1\lambda_2\lambda_3 \end{cases} \quad (251)$$

The degrees of freedom are the values at the triangle vertices and edge midpoints, together with the value at the barycenter.

Concerning the pressure approximation in the generic triangle T , we take the three shape functions

$$\begin{cases} N_1 = 1 \\ N_2 = x \\ N_3 = y \end{cases} \quad (252)$$

and the degrees of freedom can be chosen as the values at three internal and noncollinear points of the triangle.

Figure 12 displays the element degrees of freedom for both u^h and p^h .

• **The $P_{k+2} - P_k$ family.** We now present a class of mixed methods consisting in choosing, for any integer k with $k \geq 0$, the following interpolation fields.

For the approximated velocity u^h , we require that

- for each $T \in \mathcal{T}_h$, the two components of u^h are polynomials of degree at most $k+2$;

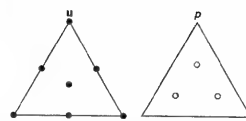


Figure 12. Degrees of freedom for Crouzeix-Raviart element.

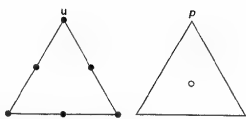
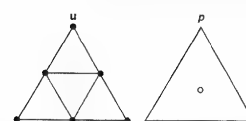
- the two components of u^h are globally continuous functions on Ω .

Instead, the approximated pressure p^h is a polynomial of degree at most k for each $T \in \mathcal{T}_h$, with no continuity imposed across the triangles.

Figure 13 shows the local degrees of freedom, for both u^h and p^h , of the lowest-order method (i.e. $k = 0$), which has been proposed and mathematically analyzed in Fortin (1975).

Remark 11. For the $P_{k+2} - P_k$ family, the discretization error in energy norm is of order h^{k+1} for both the velocity and the pressure, even though the P_{k+2} -approximation should suggest an order h^{k+2} for the velocity field. This 'suboptimality' is indeed a consequence of the poor pressure interpolation (polynomials of degrees at most k). However, taking advantage of a suitable augmented Lagrangian formulation, the $P_{k+2} - P_k$ family can be improved to obtain a convergence rate of order $h^{k+3/2}$ for the velocity, without significantly increasing the computational costs. We refer to Boffi and Lovadina (1997) for details on such an approach.

• **The $(P_1\text{-iso-}P_2) - P_0$ element.** Another stable element can be designed by taking the $P_1\text{-iso-}P_2$ element for the approximated velocity, and a piecewise constant approximation for the pressure. More precisely, as for the $(P_1\text{-iso-}P_2) - P_1^c$ element previously described, we consider a triangular mesh \mathcal{T}_h with meshsize h . We then build a finer mesh $\mathcal{T}_{h/2}$ according to the procedure sketched in Figure 10.

Figure 13. Degrees of freedom for $P_2 - P_0$ element.Figure 14. Degrees of freedom for $(P_1\text{-iso-}P_2) - P_0$ element.

We recall that the approximated velocity u^h is given using the finer mesh $\mathcal{T}_{h/2}$ and requiring that

- for each triangle of $\mathcal{T}_{h/2}$, the two components of u^h are linear functions;
- the two components of u^h are globally continuous functions on Ω .

Instead, the pressure approximation is defined on the coarser mesh \mathcal{T}_h by selecting the piecewise constant functions.

The local degrees of freedom for both u^h and p^h are shown in Figure 14.

• **The non-conforming $P_1^{NC} - P_0$ element.** We present an element, attributable to Crouzeix and Raviart (1973), for which the approximated velocity u^h is obtained by requiring that

- for each triangle the two components of u^h are linear functions;
- continuity of u^h across adjacent elements is imposed only at edge midpoints.

For the approximated pressure p^h , we simply take the piecewise constant functions.

Given a triangle $T \in \mathcal{T}_h$, the degrees of freedom for the approximating velocity u^h are the values at the three edge midpoints. Furthermore, for each component of u^h , an element basis on triangle T is provided by

$$N_k = 1 - 2\lambda_k \quad k = 1, 2, 3$$

The lack of continuity for the discrete velocity implies that the differential operators (gradient and divergence) acting on u^h should be taken element-wise. For instance, the matrix B should be written as

$$B|_T = - \sum_{i \in \mathcal{T}_h} \int_T [N_i^T \operatorname{div} (N_j^T)] d\Omega \quad (253)$$

The degrees of freedom are displayed in Figure 15. We remark that applicability of the $P_1^{NC} - P_0$ element is limited to problems with Dirichlet boundary conditions for the

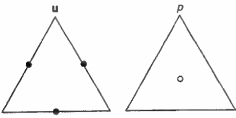


Figure 15. Degrees of freedom for $P_1^{\text{NC}} - P_0$ element.

displacement field imposed on the whole $\partial\Omega$. For other situations (e.g. a pure traction problem), the scheme exhibits spurious mechanisms, because of its inability to control the rigid body rotations (cf. Hughes, 1987). Two stable modifications of the $P_1^{\text{NC}} - P_0$ element have been proposed and analyzed in Falk (1991) and, recently, in Hansbo and Larson (2003).

4.2.3 Quadrilateral elements

Many of the triangular elements presented above have their quadrilateral counterpart. As an example, we here show the so-called $Q_2 - Q_1^c$ element, which is the quadrilateral version of the lowest-order Hood–Taylor element (cf. Figure 9). Accordingly, the velocity is approximated by biquadratic and continuous functions, while the pressure is discretized by means of bilinear and continuous functions, as depicted in Figure 16.

Another very popular scheme is the $Q_2 - P_1$ element, based on the same approximated velocities as before. Instead, the interpolating functions for the pressure are piecewise linear, without requiring any continuity across the elements. The local degrees of freedom are displayed in Figure 17.

4.2.4 Three-dimensional elements

Several elements previously described extend to the case of three-dimensional problems. In Figure 18(a), we show a continuous pressure tetrahedral element, which is nothing but the 3-D version of the MINI element (cf.

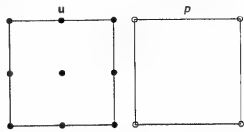


Figure 16. Degrees of freedom for $Q_2 - Q_1^c$ element.

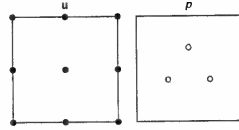


Figure 17. Degrees of freedom for $Q_2 - P_1$ element.

Figure 8). Also, the non-conforming $P_1^{\text{NC}} - P_0$ element (cf. Figure 15) has its three-dimensional counterpart, as depicted in Figure 18(b). We remark that the degrees of freedom for the velocity are given by the values at the barycenter of each tetrahedron face. Figure 18(c) shows an example of cubic element, which is exactly the 3-D version of the popular $Q_2 - P_1$ element (cf. Figure 17).

Finally, we refer to Stenberg (1987) for the analysis, based on the so-called *macroelement technique* introduced in Stenberg (1984), of the lowest-order 3-D Hood–Taylor method, and to Boffi (1997) for the higher-order case.

4.2.5 Stabilized formulations

From the above discussion, it should be clear that the fulfillment of the *inf-sup* condition requires a careful choice of the discretization spaces for the velocity and the pressure. A first strategy to obtain stability has been to derive the

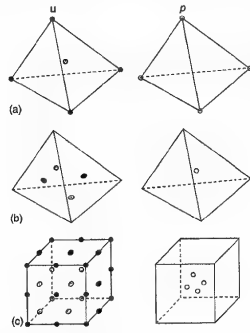


Figure 18. 3-D Stokes elements.

numerical scheme from a perturbation of functional (37), by considering (see Brezzi and Pitkäranta, 1984)

$$\begin{aligned} \tilde{L}(u, p) = & \frac{1}{2} \mu \int_{\Omega} [\nabla u : \nabla u] \, d\Omega - \int_{\Omega} [b \cdot u] \, d\Omega \\ & - \int_{\Omega} [p \operatorname{div} u] \, d\Omega - \frac{\alpha}{2} \sum_{K \in \mathcal{T}_h} \int_K h_K^2 |\nabla p|^2 \, d\Omega \end{aligned} \quad (254)$$

where α is a positive parameter and h_K is the diameter of the element $K \in \mathcal{T}_h$. The ‘perturbation term’

$$\frac{\alpha}{2} \sum_{K \in \mathcal{T}_h} \int_K h_K^2 |\nabla p|^2 \, d\Omega$$

has a stabilizing effect on the discretized Euler–Lagrange equations emanating from (254). It however introduces a consistency error, so that the convergence rate in energy norm cannot be better than $O(h)$, even though higher-order elements are used. Following the ideas in Hughes and Franca (1987) and Hughes *et al.* (1986), this drawback may be overcome by means of a suitable augmented Lagrangian formulation. Instead of considering (37) or (254), one can introduce the augmented functional

$$\begin{aligned} L^{\text{aug}}(u, p) = & \frac{1}{2} \mu \int_{\Omega} [\nabla u : \nabla u] \, d\Omega \\ & - \int_{\Omega} [b \cdot u] \, d\Omega - \int_{\Omega} [p \operatorname{div} u] \, d\Omega \\ & - \frac{1}{2} \sum_{K \in \mathcal{T}_h} \int_K \alpha(K) [\mu \Delta u - \nabla p + b]^2 \, d\Omega \end{aligned} \quad (255)$$

where, for each element $K \in \mathcal{T}_h$, $\alpha(K)$ is a positive parameter at our disposal. Because of the structure of the ‘additional term’ in (255), both the functionals (37) and (255) have the same critical point, that is, the solution of the Stokes problem. Therefore, the discretized Euler–Lagrange equations associated with (255) deliver a consistent method, whenever conforming approximations have been selected. As before, the augmented term may have a stabilizing effect, allowing the choice of a wider class of elements. For instance, if

$$\alpha(K) = \tilde{\alpha} h_K^2$$

where $\tilde{\alpha}$ is sufficiently ‘small’, any finite element approximation of velocity and pressure (as far as the pressure is discretized with continuous finite elements) leads to a stable scheme, with respect to an appropriate norm (see Franca and Hughes, 1988).

This approach has several interesting variants. Indeed, considering the Euler–Lagrange equations associated

with (255) we have

$$\begin{aligned} \mu \int_{\Omega} [\nabla u : \nabla v] \, d\Omega - \int_{\Omega} [b \cdot v] \, d\Omega - \int_{\Omega} [p \operatorname{div} v] \, d\Omega \\ - \int_{\Omega} [q \operatorname{div} u] \, d\Omega - \sum_{K \in \mathcal{T}_h} \int_K \alpha(K) [\mu \Delta u - \nabla p + b] \\ \cdot [\mu \Delta v - \nabla q] \, d\Omega = 0 \end{aligned} \quad (256)$$

for all test functions u and q . The term in second line of (256) represents our *consistent perturbation*. A careful analysis can show that its stabilizing effect still works if we change it into

$$+ \sum_{K \in \mathcal{T}_h} \int_K \alpha(K) [\mu \Delta u - \nabla p + b] \cdot [\mu \Delta v + \nabla q] \, d\Omega \quad (257)$$

(that is, changing the sign of the whole term, but changing also the sign of ∇q in the second factor) or simply into

$$+ \sum_{K \in \mathcal{T}_h} \int_K \alpha(K) [\mu \Delta u - \nabla p + b] \cdot \nabla q \, d\Omega \quad (258)$$

For a general analysis of these possible variants, we refer to Baiocchi and Brezzi (1993). A deeper analysis shows that, in particular, the formulation (257) can be interpreted as changing the space of velocities with the addition of suitable bubble functions and then eliminate them by static condensation. This was pointed out first in Pierre (1989), and then in a fully systematic way in Baiocchi, Brezzi and Franca (1993).

Other possible stabilizations can be obtained by adding penalty terms that penalize the jumps in the pressure variable over suitable macroelements. See, for instance, Silvester and Kechar (1990). This as well can be seen as adding suitable bubbles on the macroelements and eliminating them by static condensation.

For a more general survey of these and other types of stabilizations, see Brezzi and Fortin (2001) and the references therein.

Another approach to get stable elements is based on the so-called *Enhanced Strain Technique*, introduced in Simo and Rifai (1990) in the context of elasticity problems. As already mentioned in Section 2.3, the basic idea consists in enriching the symmetric gradients $\nabla^s u^h$ with additional local modes. An analysis of this strategy for displacement-based elements has been developed in Reddy and Simo (1995) and Braess (1998). Within the framework of the (u, π) formulation for incompressible elasticity problems (and therefore for the Stokes problem), the enhanced strain technique has been successfully used in Pantuso and Bathe

(1995) (see also Lovadina, 1997 for a stability and convergence analysis), and more recently in Lovadina and Auricchio (2003) and Auricchio *et al.* (submitted).

4.3 Elasticity

We now briefly consider the elasticity problem in the framework of the Hellinger–Reissner variational principle (59). We recall that after discretization we are led to solve a problem of the following type (cf. (57), (58), and (61)):

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\sigma} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{g} \end{bmatrix} \quad (259)$$

where

$$\begin{cases} \mathbf{A}_{ij} = \int_{\Omega} [\mathbf{N}_i^{\sigma} : \mathbf{D}^{-1} \mathbf{N}_j^{\sigma}] d\Omega, & \hat{\sigma}_i = \hat{\sigma}_j \\ \mathbf{B}_{lj} = \int_{\Omega} [\mathbf{N}_l^{\sigma} \cdot \text{div}(\mathbf{N}_j^{\mathbf{u}})] d\Omega, & \hat{\mathbf{u}}_l = \hat{\mathbf{u}}_r \\ \mathbf{g}_l = - \int_{\Omega} [\mathbf{N}_l^{\sigma} \cdot \mathbf{b}] d\Omega \end{cases} \quad (260)$$

Above, \mathbf{N}_i^{σ} and $\mathbf{N}_j^{\mathbf{u}}$ are the interpolation functions for the stress σ and the displacement \mathbf{u} respectively. Moreover, $\hat{\sigma}$ and $\hat{\mathbf{u}}$ are the vectors of stress and displacement unknowns. We note that since the divergence operator acts on the shape functions \mathbf{N}_i^{σ} (see the \mathbf{B} matrix in (260)), the approximated normal stress σ^h should be continuous across adjacent elements. On the contrary, no derivative operator acts on the shape functions $\mathbf{N}_i^{\mathbf{u}}$, so that we are allowed to use discontinuous approximation for the displacement field. Analogously to the thermal diffusion problem (see Section 4.1), the proper norms for $\hat{\sigma}$ and $\hat{\mathbf{u}}$ are as follows (cf. (223) and (224)):

$$\|\hat{\sigma}\|_{\mathbf{X}}^2 := \int_{\Omega} [(\mathbf{N}_i^{\sigma} \hat{\sigma}_i) : \mathbf{D}^{-1} (\mathbf{N}_j^{\sigma} \hat{\sigma}_j)] d\Omega + \frac{\ell^2}{D_*} \int_{\Omega} |\text{div}(\mathbf{N}_i^{\sigma} \hat{\sigma}_i)|^2 d\Omega \quad (261)$$

and

$$\|\hat{\mathbf{u}}\|_{\mathbf{Y}}^2 := \int_{\Omega} |\mathbf{N}_i^{\mathbf{u}} \hat{\mathbf{u}}_i|^2 d\Omega \quad (262)$$

where ℓ is some characteristic length of the domain Ω and D_* is some characteristic value of the elastic tensor.

Despite the apparent similarity with the corresponding (221) to (222) of the thermal diffusion problem, finding approximation spaces for (259) to (260), which satisfy both the *inf-sup* and the *elker* conditions, is much more difficult (see e.g. Brezzi and Fortin, 1991 for a discussion on such

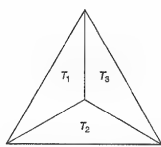


Figure 19. Splitting of a generic triangle for the Johnson–Mercier element.

a point). Here below, we present two triangular elements proposed and analyzed in Johnson and Mercier (1978) and Arnold and Winther (2002) respectively.

• **The Johnson–Mercier element.** This method takes advantage of a ‘composite’ approximation for the stress field. More precisely, we first split every triangle $T \in \mathcal{T}_h$ into three subtriangles T_i ($i = 1, 2, 3$) using the barycenter of T (see Figure 19).

For the approximated stress σ^h , we then require that

- in each subtriangle T_i the components of σ^h are linear functions;
- the normal stress $\sigma^h \mathbf{n}$ is continuous across adjacent triangles and across adjacent subtriangles.

Accordingly, the discrete stress σ^h is not a polynomial on T , but only on the subtriangles T_i . For the generic element $T \in \mathcal{T}_h$, it can be shown (see Johnson and Mercier, 1978) that the elemental degrees of freedom are the following.

- On the three edges of T : the moments of order 0 and 1 for the vector field $\sigma^h \mathbf{n}$ (12 degrees of freedom);
- On T : the moments of order 0 for the symmetric tensor field σ^h (3 degrees of freedom).

Moreover, each component of the approximated displacement \mathbf{u}^h is chosen as a piecewise constant function.

Figure 20 displays the element degrees of freedom for both σ^h and \mathbf{u}^h .

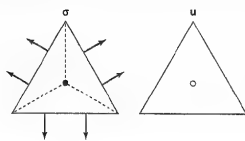


Figure 20. Degrees of freedom for the Johnson–Mercier element.

• **The Arnold–Winther element.** This triangular element has been recently proposed and analyzed in Arnold and Winther (2002), where higher-order schemes are also considered. For the approximated stress σ^h , we impose that

- on each $T \in \mathcal{T}_h$, σ^h is a symmetric tensor whose components are cubic functions, but $\text{div} \sigma^h$ is a linear vector field;
- the normal stress $\sigma^h \mathbf{n}$ is continuous across adjacent triangles.

For each element $T \in \mathcal{T}_h$, the approximation space for the stress field has dimension 24 and the elemental degrees of freedom can be chosen as follows (see Arnold and Winther, 2002):

- the values of the symmetric tensor field σ^h at the vertices of T (9 degrees of freedom);
- the moments of order 0 and 1 for the vector field $\sigma^h \mathbf{n}$ on each edge of T (12 degrees of freedom);
- the moment of order 0 for σ^h on T (3 degrees of freedom).

Furthermore, the components of the approximated displacement \mathbf{u}^h are piecewise linear functions, without requiring any continuity across adjacent elements.

In Figure 21, the element degrees of freedom for both σ^h and \mathbf{u}^h are schematically depicted.

Remark 12. Other methods exploiting ‘composite’ approximations as for the Johnson–Mercier element have been proposed and analyzed in Arnold, Douglas and Gupta (1984).

Following the ideas in Fraeijs de Veubeke (1975), a different strategy to obtain reliable schemes for the elasticity problem in the context of the Hellinger–Reissner variational principle consists in the use of unsymmetric approximated stresses. Symmetry is then enforced back in a weak form by the introduction of a suitable Lagrange multiplier. We refer to Amara and Thomas (1979), Arnold, Brezzi and Douglas (1984), Brezzi *et al.* (1986), and Stenberg (1988) for the details on such an approach.

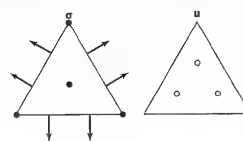


Figure 21. Degrees of freedom for the Arnold–Winther element.

5 TECHNIQUES FOR PROVING THE *INF-SUP* CONDITION

In this section we give some hints on how to prove the *inf-sup* condition (118). We also show how the stability results detailed in Section 3 can be exploited to obtain error estimates. We focus on the Stokes problem, as a representative example, but analogous strategies can be applied to analyze most of the methods considered in Section 4.

We begin recalling (cf. (38)) that a weak form of the Stokes problem with homogeneous boundary conditions for the velocity consists in finding (\mathbf{u}, p) such that

$$\begin{cases} \mu \int_{\Omega} [(\nabla \mathbf{u}) : \nabla \mathbf{u}] d\Omega - \int_{\Omega} (\text{div}(\mathbf{u})) p d\Omega \\ = \int_{\Omega} (\mathbf{b} \cdot \mathbf{u}) d\Omega \end{cases} \quad (263)$$

for any admissible velocity variation $\delta \mathbf{u}$ and any admissible pressure variation δp . On the other hand, as detailed in Section 4.2, the discretized problem consists in solving

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} \quad (264)$$

where

$$\begin{cases} \mathbf{A}_{ij} = \mu \int_{\Omega} [\nabla \mathbf{N}_i^{\mathbf{u}} : \nabla \mathbf{N}_j^{\mathbf{u}}] d\Omega, & \hat{\mathbf{u}}_i = \hat{\mathbf{u}}_j \\ \mathbf{B}_{lj} = - \int_{\Omega} [\mathbf{N}_l^{\mathbf{u}} \cdot \text{div}(\mathbf{N}_j^p)] d\Omega, & \hat{p}_l = \hat{p}_r \\ \mathbf{f}_i = \int_{\Omega} [\mathbf{N}_i^{\mathbf{u}} \cdot \mathbf{b}] d\Omega \end{cases} \quad (265)$$

with $i, j = 1, \dots, n$ and $r = 1, \dots, m$.

With our notation for the Stokes problem, the *inf-sup* condition in its equivalent form (119) consists in requiring the existence of a positive constant β , independent of h , such that

$$\forall \hat{\mathbf{q}} \in \mathbf{Y} \quad \sup_{\hat{\mathbf{z}} \in \mathbf{X}(0)} \frac{\hat{\mathbf{z}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{z}}\|_{\mathbf{X}}} \geq \beta \|\hat{\mathbf{q}}\|_{\mathbf{Y}} \quad (266)$$

where $\mathbf{X} = \mathbb{R}^n$ and $\mathbf{Y} = \mathbb{R}^m$.

Moreover, in what follows, we need to introduce the space \mathcal{X} for vectorial functions \mathbf{v} , defined by

$$\mathcal{X} = \left\{ \mathbf{v} : \mathbf{v}|_{\partial\Omega} = \mathbf{0}, \quad \|\mathbf{v}\|_{\mathcal{X}}^2 := \mu \int_{\Omega} |\nabla \mathbf{v}|^2 d\Omega < +\infty \right\} \quad (267)$$

and the space \mathcal{V} for scalar functions q , defined by

$$\mathcal{V} = \left\{ q : \|q\|_{\mathcal{V}}^2 := \int_{\Omega} |q|^2 d\Omega < +\infty \right\} \quad (268)$$

Remark 13. It is worth noticing that, whenever an approximated velocity $\mathbf{u}^h = \mathbf{N}_i^h \hat{u}_i$ is considered, the following holds (cf. (243))

$$\|\hat{\mathbf{u}}\|_X = \left(\mu \int_{\Omega} |\nabla \mathbf{N}_i^h \hat{u}_i|^2 d\Omega \right)^{1/2} = \|\mathbf{u}^h\|_X \quad (269)$$

Therefore, the X -norm we have tailored for vector $\hat{\mathbf{u}} \in \mathbf{X}$ coincides with the X -norm of the reconstructed function \mathbf{u}^h . Similarly (cf. (244)), if $p^h = N^p \hat{p}$, we have

$$\|\hat{p}\|_Y = \left(\mu \int_{\Omega} |N^p \hat{p}|^2 d\Omega \right)^{1/2} = \|p^h\|_Y \quad (270)$$

5.1 Checking the inf-sup condition

As already mentioned, a rigorous proof of the inf-sup condition is typically a difficult task, mainly because several technical mathematical problems have to be overcome. In this section, we present two of the most powerful tools for proving the inf-sup property. The first technique (*Fortin's trick*) can be used, for instance, to study the stability of the $P_1^{\text{NC}} - P_0$ element (cf. Figure 15) and the Crouzeix–Raviart element (cf. Figure 12), as we are going to detail below. The second one (*Verfürth's trick*) can be applied basically to all the approximations with continuous pressure and it will be exemplified by considering the MINI element (cf. Figure 8).

Although we are aware that the subsequent analysis is not completely satisfactory from the mathematical point of view, it nonetheless highlights some of the basic ideas behind the analysis of mixed finite element methods.

We first need to recall the following important theorem of functional analysis (see Ladyzhenskaya, 1969; Temam, 1977, for instance).

Theorem 4. *There exists a constant $\beta_c > 0$ such that, for every $q \in \mathcal{V}$ with $\int_{\Omega} q d\Omega = 0$, it holds*

$$\sup_{\mathbf{v} \in \mathbf{X} \setminus \{0\}} \frac{-\int_{\Omega} q \operatorname{div} \mathbf{v} d\Omega}{\|\mathbf{v}\|_X} \geq \beta_c \|q\|_Y \quad (271)$$

Remark 14. We remark that estimate (271) is nothing but the infinite-dimensional version of the inf-sup condition written in its equivalent form (119).

5.1.1 Fortin's trick

The next result provides a criterion for proving the inf-sup condition, called *Fortin's trick* (see Fortin, 1977) or, more precisely, Fortin's trick applied to the Stokes problem.

Proposition 4. *Suppose there exists a linear operator $\hat{\Pi}_h : \mathcal{X} \rightarrow \mathbf{X} \equiv \mathbb{R}^n$ such that*

$$\|\hat{\Pi}_h \mathbf{v}\|_X \leq C_{\hat{\Pi}} \|\mathbf{v}\|_X \quad \forall \mathbf{v} \in \mathcal{X} \quad (272)$$

and

$$(\hat{\Pi}_h \mathbf{v})^T \mathbf{B}^T \hat{\mathbf{q}} = - \int_{\Omega} \operatorname{div} \mathbf{v} (N^p \hat{q}_r) d\Omega \quad \forall \hat{\mathbf{q}} \in \mathbf{Y} \equiv \mathbb{R}^m \quad (273)$$

with $C_{\hat{\Pi}}$ independent of h . Then it holds

$$\forall \hat{\mathbf{q}} \in \mathbf{Y} \quad \sup_{\mathbf{v} \in \mathcal{X} \setminus \{0\}} \frac{\hat{\mathbf{v}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\mathbf{v}\|_X} \geq \frac{\beta_c}{C_{\hat{\Pi}}} \|\hat{\mathbf{q}}\|_Y \quad (274)$$

that is, the inf-sup condition (266) is fulfilled with $\beta = \beta_c / C_{\hat{\Pi}}$.

Proof. Take any $\hat{\mathbf{q}} \in \mathbf{Y}$. We notice that from Theorem 4 and Remark 13, we get

$$\sup_{\mathbf{v} \in \mathcal{X} \setminus \{0\}} \frac{-\int_{\Omega} \operatorname{div} \mathbf{v} (N^p \hat{q}_r) d\Omega}{\|\mathbf{v}\|_X} \geq \beta_c \|N^p \hat{q}_r\|_Y = \beta_c \|\hat{\mathbf{q}}\|_Y \quad (275)$$

Therefore, from (273), we have

$$\sup_{\mathbf{v} \in \mathcal{X} \setminus \{0\}} \frac{(\hat{\Pi}_h \mathbf{v})^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\mathbf{v}\|_X} \geq \beta_c \|\hat{\mathbf{q}}\|_Y \quad (276)$$

Using (272), from (276) it follows

$$\begin{aligned} \sup_{\mathbf{v} \in \mathcal{X} \setminus \{0\}} \frac{(\hat{\Pi}_h \mathbf{v})^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\Pi}_h \mathbf{v}\|_X} &\geq \frac{1}{C_{\hat{\Pi}}} \sup_{\mathbf{v} \in \mathcal{X} \setminus \{0\}} \frac{(\hat{\Pi}_h \mathbf{v})^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\mathbf{v}\|_X} \\ &\geq \frac{\beta_c}{C_{\hat{\Pi}}} \|\hat{\mathbf{q}}\|_Y \end{aligned} \quad (277)$$

Since, obviously, $\{\hat{\Pi}_h \mathbf{v} : \mathbf{v} \in \mathcal{X}\} \subseteq \mathbf{X}$, from (277) we obtain

$$\sup_{\hat{\mathbf{v}} \in \mathbf{X} \setminus \{0\}} \frac{\hat{\mathbf{v}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{v}}\|_X} \geq \sup_{\mathbf{v} \in \mathcal{X} \setminus \{0\}} \frac{(\hat{\Pi}_h \mathbf{v})^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\Pi}_h \mathbf{v}\|_X} \geq \frac{\beta_c}{C_{\hat{\Pi}}} \|\hat{\mathbf{q}}\|_Y \quad \square \quad (278)$$

We now apply Proposition 4 to the $P_1^{\text{NC}} - P_0$ element and the Crouzeix–Raviart element, skipping, however, the proof of (272). In both cases, the strategy for building the operator $\hat{\Pi}_h$ is the following:

1. On each triangle $T \in \mathcal{T}_h$, we first define a suitable linear operator $\Pi_{h,T} : \mathbf{V} \mapsto \Pi_{h,T} \mathbf{v}$, valued in the space of velocity approximating functions on T , and satisfying

$$\int_T q^h \operatorname{div} (\Pi_{h,T} \mathbf{v}) d\Omega = \int_T q^h \operatorname{div} \mathbf{v} d\Omega \quad (279)$$

for every $\mathbf{v} \in \mathcal{X}$ and every discrete pressure q^h . This will be done by using, in particular, the element degrees of freedom for the velocity approximation.

2. By assembling all the element contributions, we obtain a global linear operator

$$\Pi_h : \begin{cases} \mathcal{X} \longrightarrow \operatorname{Span}\{\mathbf{N}_i^h; i = 1, \dots, n\} \\ \mathbf{v} \longmapsto \Pi_h \mathbf{v} = \sum_{T \in \mathcal{T}_h} \Pi_{h,T} \mathbf{v} = \mathbf{N}_i^h \hat{v}_i \end{cases} \quad (280)$$

3. We finally define $\hat{\Pi}_h : \mathcal{X} \rightarrow \mathbb{R}^n$ by setting

$$\hat{\Pi}_h \mathbf{v} = \hat{\mathbf{v}} \quad \text{if} \quad \Pi_h \mathbf{v} = \mathbf{N}_i^h \hat{v}_i \quad (281)$$

that is, $\hat{\Pi}_h \mathbf{v}$ returns the components of the function $\Pi_h \mathbf{v}$ with respect to the global velocity basis $\{\mathbf{N}_i^h; i = 1, \dots, n\}$. From the definition of the matrix \mathbf{B} , property (279), (280), and (281), it follows that condition (273) is satisfied.

- *The $P_1^{\text{NC}} - P_0$ element.* Fix $T \in \mathcal{T}_h$, and recall that any approximated pressure q^h is a constant function on T . We wish to build $\Pi_{h,T}$ in such a way that

$$\int_T q^h \operatorname{div} (\Pi_{h,T} \mathbf{v}) d\Omega = \int_T q^h \operatorname{div} \mathbf{v} d\Omega \quad (282)$$

From the divergence theorem, (282) can be alternatively written as

$$\int_{\partial T} q^h (\Pi_{h,T} \mathbf{v}) \cdot \mathbf{n} ds = \int_{\partial T} q^h \mathbf{v} \cdot \mathbf{n} ds \quad (283)$$

Denoting with M_k ($k = 1, 2, 3$) the midpoint of the edge e_k , we define $\Pi_{h,T} \mathbf{v}$ as the unique (vectorial) linear function such that

$$\Pi_{h,T} \mathbf{v}(M_k) = \frac{1}{|e_k|} \int_{e_k} \mathbf{v} ds \quad k = 1, 2, 3 \quad (284)$$

From the divergence theorem and the Midpoint rule, it follows that

$$\begin{aligned} \int_T q^h \operatorname{div} (\Pi_{h,T} \mathbf{v}) d\Omega &= \int_{\partial T} q^h (\Pi_{h,T} \mathbf{v}) \cdot \mathbf{n} ds \\ &= \int_{\partial T} q^h \mathbf{v} \cdot \mathbf{n} ds = \int_T q^h \operatorname{div} \mathbf{v} d\Omega \end{aligned} \quad (285)$$

for every constant function q^h . It is now sufficient to define the global linear operator Π_h as

$$\Pi_h \mathbf{v} = \sum_{T \in \mathcal{T}_h} \Pi_{h,T} \mathbf{v} = \mathbf{N}_i^h \hat{v}_i$$

and the corresponding operator $\hat{\Pi}_h$ satisfies condition (273) (cf. also (253)).

- *The Crouzeix–Raviart element.* Fix $T \in \mathcal{T}_h$, and recall that any approximated pressure q^h is now a linear function on T . Hence, q^h can be uniquely decomposed as $q^h = q_0 + q_1$, where q_0 is a constant (the mean value of q^h on T), and q_1 is a linear function having zero mean value. We now construct a linear operator $\Pi_{1,T} : \mathbf{V} \mapsto \Pi_{1,T} \mathbf{v}$, where $\Pi_{1,T} \mathbf{v}$ is a quadratic vectorial polynomial such that

$$\int_T \operatorname{div} (\Pi_{1,T} \mathbf{v}) d\Omega = \int_T \operatorname{div} \mathbf{v} d\Omega \quad (286)$$

or, alternatively,

$$\int_{\partial T} (\Pi_{1,T} \mathbf{v}) \cdot \mathbf{n} ds = \int_{\partial T} \mathbf{v} \cdot \mathbf{n} ds \quad (287)$$

Denoting with V_k (resp., M_k) the vertexes of T (resp., the midpoint of the edge e_k), the Cavalieri–Simpson rule shows that condition (287) holds if we set

$$\begin{cases} \Pi_{1,T} \mathbf{v}(V_k) = \mathbf{v}(V_k) & k = 1, 2, 3 \\ \Pi_{1,T} \mathbf{v}(M_k) = \frac{3}{2|e_k|} \int_{e_k} \mathbf{v} ds - \frac{\mathbf{v}(V_{k_1}) + \mathbf{v}(V_{k_2})}{4} & k = 1, 2, 3 \end{cases} \quad (288)$$

Above, we have denoted with V_{k_1} and V_{k_2} the endpoints of side e_k . So far, we have not used the bubble functions available for the approximated velocity. We now use these two additional degrees of freedom by defining $\mathbf{v}_{b,T}(\mathbf{v})$ as the unique vectorial bubble function such that

$$\int_T q_1 \operatorname{div} \mathbf{v}_{b,T}(\mathbf{v}) d\Omega = \int_T q_1 \operatorname{div} (\mathbf{v} - \Pi_{1,T} \mathbf{v}) d\Omega \quad (289)$$

for every linear function q_1 having zero mean value on T .

We claim that if $\Pi_{h,T} \mathbf{v} = \mathbf{v}_{b,T}(\mathbf{v}) + \Pi_{1,T} \mathbf{v}$, then

$$\int_T q^h \operatorname{div} (\Pi_{h,T} \mathbf{v}) d\Omega = \int_T q^h \operatorname{div} \mathbf{v} d\Omega \quad (290)$$

for every linear polynomial $q^h = q_0 + q_1$. In fact, using (286), (289), and the obvious fact that $\int_T \operatorname{div} \mathbf{v}_{b,T}(\mathbf{v}) d\Omega = 0$,

$d\Omega = 0$, we have

$$\begin{aligned} \int_T q^h \operatorname{div}(\Pi_{h,T} \mathbf{v}) \, d\Omega &= \int_T (q_0 + q_1) \operatorname{div}(\mathbf{v}_{b,T}(\mathbf{v})) \\ &\quad + \Pi_{1,T} \mathbf{v}) \, d\Omega = \int_T q_0 \operatorname{div}(\mathbf{v}_{b,T}(\mathbf{v}) + \Pi_{1,T} \mathbf{v}) \, d\Omega \\ &\quad + \int_T q_1 \operatorname{div}(\mathbf{v}_{b,T}(\mathbf{v}) + \Pi_{1,T} \mathbf{v}) \, d\Omega \\ &= \int_T q_0 \operatorname{div}(\Pi_{1,T} \mathbf{v}) \, d\Omega + \int_T q_1 \operatorname{div} \mathbf{v} \, d\Omega \\ &= \int_T q_0 \operatorname{div} \mathbf{v} \, d\Omega + \int_T q_1 \operatorname{div} \mathbf{v} \, d\Omega \\ &= \int_T q^h \operatorname{div} \mathbf{v} \, d\Omega \end{aligned} \quad (291)$$

Hence, the operator $\hat{\Pi}_h$ arising from the global linear operator

$$\Pi_h \mathbf{v} = \sum_{T \in \mathcal{T}_h} (\mathbf{v}_{b,T}(\mathbf{v}) + \Pi_{1,T} \mathbf{v}) = \mathbf{v}_b(\mathbf{v}) + \Pi_1 \mathbf{v} = \mathbf{N}_1^0 \hat{\mathbf{v}},$$

fulfills condition (273).

Remark 15. Conditions (288) reveal that the operator Π_1 (built by means of the local contributions $\Pi_{1,T}$) exploits, in particular, the point values of \mathbf{v} at all the vertices of the triangles in \mathcal{T}_h . However, this definition makes no sense for an arbitrary $\mathbf{v} \in \mathcal{X}$, since functions in \mathcal{X} are not necessarily continuous. To overcome this problem, one should define a more sophisticated operator Π_1 for instance, taking advantage of an averaging procedure. More precisely, one could define the function $\Pi_1 \mathbf{v}$ as the unique piecewise quadratic polynomial such that $\Pi_1 \mathbf{v}|_{\partial\Omega} = 0$ and

$$\begin{cases} \Pi_1 \mathbf{v}(V) = \frac{1}{\operatorname{Area}(D(V))} \int_{D(V)} \mathbf{v} \, d\Omega \\ \Pi_1 \mathbf{v}(M) = \frac{3}{2|e_M|} \int_{e_M} \mathbf{v} \, ds - \frac{\Pi_1 \mathbf{v}(V_{M_1}) + \Pi_1 \mathbf{v}(V_{M_2})}{4} \end{cases} \quad (292)$$

Above, V is any internal vertex of triangles in \mathcal{T}_h and $D(V)$ is the union of the triangles having V as a vertex. Moreover, e_M is any internal edge having M as midpoint and V_{M_1}, V_{M_2} as endpoints. In this case, it is possible to prove that for the resulting $\hat{\Pi}_h$, the very important property (272) holds with $C_{\hat{\Pi}}$ independent of h .

Remark 16. It is interesting to observe that condition (283) and (287) suggest the following important fact about the discretization of the Stokes problem. Any reasonable discontinuous pressure approximation contains at least all the piecewise constant functions: relations (283)

and (287) show that having some velocity degrees of freedom associated with the triangle edges greatly helps in proving the *inf-sup* condition.

5.1.2 Verfürth's trick

We now describe another technique for proving the *inf-sup* condition, which can be profitably used when elements with continuous pressure interpolation are considered: the so-called *Verfürth's trick* (see Verfürth, 1984). We begin by noting that, because of the pressure continuity, it holds

$$\begin{aligned} \hat{\mathbf{z}}^T \mathbf{B}^T \hat{\mathbf{q}} &= - \int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot \operatorname{div}(\mathbf{N}_1^0 \hat{\mathbf{z}}_r) \, d\Omega \\ &= \int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot \mathbf{N}_1^0 \hat{\mathbf{z}}_r \, d\Omega \end{aligned} \quad (293)$$

for every $\hat{\mathbf{z}} \in \mathcal{X}$ and $\hat{\mathbf{q}} \in \mathcal{Y}$. In some cases, it is much easier to use the form (293) and prove a modified version of the *inf-sup* condition (266) with a norm for \mathcal{Y} different from the one defined in (270) and involving the pressure gradients; see Bercovier and Pironneau (1977) and Glowinski and Pironneau (1979). More precisely, given a mesh \mathcal{T}_h , we introduce in \mathcal{Y} the norm

$$\|\hat{\mathbf{p}}\|_{Y_h} = \left(\sum_{K \in \mathcal{T}_h} h_K^2 \int_K |\nabla N^p \hat{\mathbf{p}}_r|^2 \, d\Omega \right)^{1/2} \quad (294)$$

where h_K denotes the diameter of the generic element K . The key point of Verfürth's trick is a smart use of the properties of interpolation operator in order to prove that the *inf-sup* condition with the norm Y_h implies the usual one. Indeed, we have the following result.

Proposition 5. Suppose that

(H1) for every velocity $\mathbf{v} \in \mathcal{X}$ there exists a discrete velocity $\mathbf{v}_I = \mathbf{N}_1^0 \hat{\mathbf{v}}_I$ such that

$$\left(\sum_{K \in \mathcal{T}_h} h_K^{-2} \int_K |\mathbf{N}_1^0 \hat{\mathbf{v}}_I - \mathbf{v}|^2 \, d\Omega \right)^{1/2} \leq c_0 \|\mathbf{v}\|_{\mathcal{X}} \quad (295)$$

$$\|\hat{\mathbf{v}}_I\|_{\mathcal{X}} \leq c_1 \|\mathbf{v}\|_{\mathcal{X}} \quad (296)$$

with c_0, c_1 independent of h and \mathbf{v} ;

(H2) there exists a constant $\beta_* > 0$ independent of h such that

$$\forall \hat{\mathbf{q}} \in \mathcal{Y} \quad \sup_{\hat{\mathbf{z}} \in \mathcal{X}(\hat{\mathbf{q}})} \frac{\hat{\mathbf{z}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{z}}\|_{\mathcal{X}}} \geq \beta_* \|\hat{\mathbf{q}}\|_{Y_h} \quad (297)$$

(i.e. the *inf-sup* condition holds with the modified Y -norm (294)).

Then the *inf-sup* condition (with respect to the original norm (270))

$$\forall \hat{\mathbf{q}} \in \mathcal{Y} \quad \sup_{\hat{\mathbf{z}} \in \mathcal{X}(\hat{\mathbf{q}})} \frac{\hat{\mathbf{z}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{z}}\|_{\mathcal{X}}} \geq \beta \|\hat{\mathbf{q}}\|_{\mathcal{Y}} \quad (298)$$

is satisfied with β independent of h .

Proof. Given $\hat{\mathbf{q}} \in \mathcal{Y}$, we observe that using (296) and (293) it holds

$$\begin{aligned} \sup_{\hat{\mathbf{z}} \in \mathcal{X}(\hat{\mathbf{q}})} \frac{\hat{\mathbf{z}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{z}}\|_{\mathcal{X}}} &\geq \sup_{\hat{\mathbf{z}} \in \mathcal{X}(\hat{\mathbf{q}})} \frac{\hat{\mathbf{z}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{z}}\|_{\mathcal{X}}} \geq \sup_{\hat{\mathbf{z}} \in \mathcal{X}(\hat{\mathbf{q}})} \frac{\hat{\mathbf{z}}^T \mathbf{B}^T \hat{\mathbf{q}}}{c_1 \|\mathbf{v}\|_{\mathcal{X}}} \\ &= \sup_{\hat{\mathbf{z}} \in \mathcal{X}(\hat{\mathbf{q}})} \frac{\int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot \mathbf{N}_1^0 \hat{\mathbf{z}}_r \, d\Omega}{c_1 \|\mathbf{v}\|_{\mathcal{X}}} \end{aligned} \quad (299)$$

Furthermore, from Theorem 4, there exists $\mathbf{w} \in \mathcal{X}$ such that

$$\frac{- \int_{\Omega} \operatorname{div} \mathbf{w} (\nabla N^p \hat{\mathbf{q}}_r) \, d\Omega}{\|\mathbf{w}\|_{\mathcal{X}}} = \frac{\int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot \mathbf{w} \, d\Omega}{\|\mathbf{w}\|_{\mathcal{X}}} \geq \beta_c \|\hat{\mathbf{q}}\|_{\mathcal{Y}} \quad (300)$$

For such a velocity \mathbf{w} and the corresponding discrete velocity $\mathbf{w}_I = \mathbf{N}_1^0 \hat{\mathbf{w}}_I$, we obviously have

$$\sup_{\hat{\mathbf{z}} \in \mathcal{X}(\hat{\mathbf{q}})} \frac{\int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot \mathbf{N}_1^0 \hat{\mathbf{z}}_r \, d\Omega}{c_1 \|\mathbf{v}\|_{\mathcal{X}}} \geq \frac{\int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot \mathbf{N}_1^0 \hat{\mathbf{w}}_I \, d\Omega}{c_1 \|\mathbf{w}\|_{\mathcal{X}}} \quad (301)$$

Subtracting and adding \mathbf{w} , we obtain

$$\begin{aligned} \frac{\int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot \mathbf{N}_1^0 \hat{\mathbf{w}}_I \, d\Omega}{c_1 \|\mathbf{w}\|_{\mathcal{X}}} &= \frac{\int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot (\mathbf{N}_1^0 \hat{\mathbf{w}}_I - \mathbf{w}) \, d\Omega}{c_1 \|\mathbf{w}\|_{\mathcal{X}}} \\ &\quad + \frac{\int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot \mathbf{w} \, d\Omega}{c_1 \|\mathbf{w}\|_{\mathcal{X}}} \end{aligned} \quad (302)$$

To treat the first term in the right-hand side of (302), we observe that using (295) and recalling (294), we have

$$\begin{aligned} - \int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot (\mathbf{N}_1^0 \hat{\mathbf{w}}_I - \mathbf{w}) \, d\Omega &= - \sum_{K \in \mathcal{T}_h} \int_K (\nabla N^p \hat{\mathbf{q}}_r) \cdot (\mathbf{N}_1^0 \hat{\mathbf{w}}_I - \mathbf{w}) \, d\Omega \\ &= - \sum_{K \in \mathcal{T}_h} \int_K h_K (\nabla N^p \hat{\mathbf{q}}_r) \cdot h_K^{-1} (\mathbf{N}_1^0 \hat{\mathbf{w}}_I - \mathbf{w}) \, d\Omega \end{aligned}$$

$$\begin{aligned} &\leq \left(\sum_{K \in \mathcal{T}_h} h_K^2 \int_K |\nabla N^p \hat{\mathbf{q}}_r|^2 \, d\Omega \right)^{1/2} \\ &\quad \times \left(\sum_{K \in \mathcal{T}_h} h_K^{-2} \int_K |\mathbf{N}_1^0 \hat{\mathbf{w}}_I - \mathbf{w}|^2 \, d\Omega \right)^{1/2} \\ &\leq c_0 \|\hat{\mathbf{q}}\|_{Y_h} \|\mathbf{w}\|_{\mathcal{X}} \end{aligned} \quad (303)$$

which gives

$$\int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot (\mathbf{N}_1^0 \hat{\mathbf{w}}_I - \mathbf{w}) \, d\Omega \geq -c_0 \|\hat{\mathbf{q}}\|_{Y_h} \|\mathbf{w}\|_{\mathcal{X}} \quad (304)$$

Therefore, we get

$$\frac{\int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot (\mathbf{N}_1^0 \hat{\mathbf{w}}_I - \mathbf{w}) \, d\Omega}{c_1 \|\mathbf{w}\|_{\mathcal{X}}} \geq -\frac{c_0}{c_1} \|\hat{\mathbf{q}}\|_{Y_h} \quad (305)$$

For the second term in the right-hand side of (302), we notice that (cf. (300))

$$\frac{\int_{\Omega} (\nabla N^p \hat{\mathbf{q}}_r) \cdot \mathbf{w} \, d\Omega}{c_1 \|\mathbf{w}\|_{\mathcal{X}}} \geq \frac{\beta_c}{c_1} \|\hat{\mathbf{q}}\|_{Y_h} \quad (306)$$

Therefore, from (299), (301), (302), (305), and (306), we obtain

$$\sup_{\hat{\mathbf{z}} \in \mathcal{X}(\hat{\mathbf{q}})} \frac{\hat{\mathbf{z}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{z}}\|_{\mathcal{X}}} \geq \frac{\beta_c}{c_1} \|\hat{\mathbf{q}}\|_{Y_h} - \frac{c_0}{c_1} \|\hat{\mathbf{q}}\|_{Y_h} \quad (307)$$

We now multiply the modified *inf-sup* condition (297) by $c_0/(\beta_* c_1)$ to get

$$\frac{c_0}{\beta_* c_1} \sup_{\hat{\mathbf{z}} \in \mathcal{X}(\hat{\mathbf{q}})} \frac{\hat{\mathbf{z}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{z}}\|_{\mathcal{X}}} \geq \frac{c_0}{c_1} \|\hat{\mathbf{q}}\|_{Y_h} \quad (308)$$

By adding (307) and (308), we finally have

$$\left(1 + \frac{c_0}{\beta_* c_1} \right) \sup_{\hat{\mathbf{z}} \in \mathcal{X}(\hat{\mathbf{q}})} \frac{\hat{\mathbf{z}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{z}}\|_{\mathcal{X}}} \geq \frac{\beta_c}{c_1} \|\hat{\mathbf{q}}\|_{Y_h} \quad (309)$$

that is, the *inf-sup* condition (298) holds with $\beta = (\beta_c/c_1) (1 + (c_0/\beta_* c_1))^{-1}$. \square

Remark 17. We notice that hypothesis (H1) of Proposition 5 is not very restrictive. Indeed, given a velocity $\mathbf{v} \in \mathcal{X}$, the corresponding \mathbf{v}_I can be chosen as a suitable discrete velocity interpolating \mathbf{v} , and (295) and (296) are both satisfied basically for every element of practical interest (see e.g. Brezzi and Fortin, 1991 and Ciarlet, 1978 for more details).

The Verfürth trick was originally applied to the Hood-Taylor element depicted in Figure 9 (see Verfürth, 1984), but it was soon recognized as a valuable instrument for analyzing all continuous pressure elements. Here we show how to use it for the analysis of the MINI element (whose original proof was given using Fortin's trick in Arnold, Brezzi and Fortin, 1984).

• *The MINI element.* We now give a hint on how to verify hypothesis (H2) of Proposition 5 for the MINI element (cf. Figure 8).

For a generic $\hat{q} \in Y$, we take its reconstructed discrete pressure $q^h = N^p \hat{q}_r$. Since q^h is a piecewise linear and continuous function, it follows that $\nabla q^h = \nabla N^p \hat{q}_r$ is a well-defined piecewise constant vector field. We now construct a discrete (bubble-type) velocity $\mathbf{v}^h = N^u \hat{v}_r$, defined on each triangle $T \in \mathcal{T}_h$ as

$$\mathbf{v}^h = h_T^2 b_T \nabla q^h \quad (310)$$

where b_T is the usual cubic bubble (i.e. in area coordinates, $b_T = 27\lambda_1\lambda_2\lambda_3$). Recalling (293) and using (310), we then obtain

$$\begin{aligned} \hat{\mathbf{v}}^T \mathbf{B}^T \hat{\mathbf{q}} &= \int_{\Omega} (\nabla N^p \hat{q}_r) \cdot N^u \hat{v}_r \, d\Omega = \int_{\Omega} \nabla q^h \cdot \mathbf{v}^h \, d\Omega \\ &= \sum_{T \in \mathcal{T}_h} h_T^2 \int_T |\nabla N^p \hat{q}_r|^2 b_T \, d\Omega \end{aligned} \quad (311)$$

It is easy to show that for regular meshes (roughly: for meshes that do not contain 'too thin' elements, see e.g. Ciarlet, 1978 for a precise definition), there exists a constant $C_1 > 0$, independent of h , such that

$$\forall T \in \mathcal{T}_h \quad \int_T |\nabla N^p \hat{q}_r|^2 b_T \, d\Omega \geq C_1 \int_T |\nabla N^p \hat{q}_r|^2 \, d\Omega \quad (312)$$

Therefore, from (311), (312), and (294) we get

$$\hat{\mathbf{v}}^T \mathbf{B}^T \hat{\mathbf{q}} \geq C_1 \sum_{T \in \mathcal{T}_h} h_T^2 \int_T |\nabla N^p \hat{q}_r|^2 \, d\Omega = C_1 \|\hat{\mathbf{q}}\|_{Y_h}^2 \quad (313)$$

Furthermore, using standard scaling arguments (cf. Brezzi and Fortin, 1991), it is possible to prove that there exists $C_2 > 0$ independent of h such that

$$\begin{aligned} \|\hat{\mathbf{v}}\|_X &= \|\mathbf{v}^h\|_X \leq C_2 \left(\sum_{T \in \mathcal{T}_h} h_T^2 \int_T |\nabla N^p \hat{q}_r|^2 \, d\Omega \right)^{1/2} \\ &= C_2 \|\hat{\mathbf{q}}\|_{Y_h} \end{aligned} \quad (314)$$

Hence, estimates (313) and (314) imply

$$\frac{\hat{\mathbf{v}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{v}}\|_X} \geq \frac{C_1}{C_2} \|\hat{\mathbf{q}}\|_{Y_h} \quad (315)$$

and condition (297) then follows with $\beta_s = C_1/C_2$, since

$$\sup_{\hat{\mathbf{q}} \in X \setminus \{0\}} \frac{\hat{\mathbf{v}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{q}}\|_X} \geq \frac{\hat{\mathbf{v}}^T \mathbf{B}^T \hat{\mathbf{q}}}{\|\hat{\mathbf{v}}\|_X} \quad (316)$$

5.2 Appendix — error estimates

In this brief Appendix, we present the guidelines to obtain error estimates, once the stability conditions have been established. We only consider the easiest case of conforming schemes (i.e. when the velocity is approximated by means of continuous functions). We refer to Brezzi (1974), Brezzi and Fortin (1991), and Braess (1997) for more details, as well as for the analysis of more complicated situations involving non-conforming approximations (such as the $P_1^{\text{NC}} - P_0$ element (cf. Figure 15)).

Before proceeding, we recall that for the Stokes problem with our choices of norms, we have $M_u = 1$, $M_p = \sqrt{(d/\mu)}$, and $\alpha = 1$, no matter what the approximations of velocity and pressure are. However, in the subsequent discussion, we will not substitute these values into the estimates, in order to facilitate the extension of the analysis to other problems. We also notice that, on the contrary, the relevant constant β does depend on the choice of the interpolating functions. We have the following result.

Theorem 5. Let (\mathbf{u}, p) be the solution of problem (263) and suppose there exist discrete velocity and pressure

$$\mathbf{u}_I = N^u \hat{\mathbf{u}}_I, \quad p_I = N^p \hat{p}_I \quad (317)$$

such that

$$\|\mathbf{u} - \mathbf{u}_I\|_X \leq Ch^k, \quad k_u > 0 \quad (318)$$

$$\|p - p_I\|_Y \leq Ch^{k_p}, \quad k_p > 0 \quad (319)$$

If $(\hat{\mathbf{u}}, \hat{p})$ is the solution of the discrete problem (264), then, setting $\mathbf{u}^h = N^u \hat{\mathbf{u}}_I$ and $p^h = N^p \hat{p}_I$, it holds

$$\|\mathbf{u} - \mathbf{u}^h\|_X + \|p - p^h\|_Y \leq Ch^k \quad (320)$$

with $k = \min\{k_u, k_p\}$.

Proof. For \mathbf{u}_I and p_I as in (317), we set $\hat{\mathbf{u}}_I = (\hat{\mathbf{u}}_I^r)_{r=1}^n \in X$ and $\hat{p}_I = (\hat{p}_I^r)_{r=1}^m \in Y$. Taking into account that

$$\begin{Bmatrix} \hat{\mathbf{u}} \\ \hat{p} \end{Bmatrix}$$

is the solution of the discretized problem (264), we obtain that

$$\begin{Bmatrix} \hat{\mathbf{u}} - \hat{\mathbf{u}}_I \\ \hat{p} - \hat{p}_I \end{Bmatrix}$$

solves

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & 0 \end{bmatrix} \begin{Bmatrix} \hat{\mathbf{u}} - \hat{\mathbf{u}}_I \\ \hat{p} - \hat{p}_I \end{Bmatrix} = \begin{Bmatrix} \mathbf{f} - \mathbf{A}\hat{\mathbf{u}}_I - \mathbf{B}^T\hat{p}_I \\ -\mathbf{B}\hat{\mathbf{u}}_I \end{Bmatrix} \quad (321)$$

Choosing as (admissible) velocity and pressure variations the interpolating shape functions, from (263), we have

$$\begin{aligned} \mathbf{f}_I &= \mu \int_{\Omega} \nabla N_i^u : \nabla \mathbf{u} \, d\Omega - \int_{\Omega} \text{div} (N_i^u) p \, d\Omega \\ i &= 1, \dots, n \end{aligned} \quad (322)$$

and

$$\int_{\Omega} N_s^p \text{div} \mathbf{u} \, d\Omega = 0 \quad s = 1, \dots, m \quad (323)$$

Hence, system (321) may be written as

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & 0 \end{bmatrix} \begin{Bmatrix} \hat{\mathbf{u}} - \hat{\mathbf{u}}_I \\ \hat{p} - \hat{p}_I \end{Bmatrix} = \begin{Bmatrix} \tilde{\mathbf{f}} \\ \tilde{g} \end{Bmatrix} \quad (324)$$

where

$$\begin{aligned} \tilde{\mathbf{f}}_i &:= \mu \int_{\Omega} \nabla N_i^u : \nabla (\mathbf{u} - \mathbf{u}_I) \, d\Omega \\ &\quad - \int_{\Omega} \text{div} (N_i^u) (p - p_I) \, d\Omega \quad i = 1, \dots, n \end{aligned} \quad (325)$$

and

$$\tilde{g}_r := - \int_{\Omega} N_r^p \text{div} (\mathbf{u} - \mathbf{u}_I) \, d\Omega \quad r = 1, \dots, m \quad (326)$$

Applying Theorem 1, we thus obtain

$$\|\hat{\mathbf{u}} - \hat{\mathbf{u}}_I\|_X \leq \frac{1}{\alpha} \|\tilde{\mathbf{f}}\|_F + \frac{M_u^{1/2}}{\alpha^{1/2}\beta} \|\tilde{g}\|_G \quad (327)$$

$$\|\hat{p} - \hat{p}_I\|_Y \leq \left(\frac{1}{\beta} + \frac{M_p^{1/2}}{\alpha^{1/2}\beta} \right) \|\tilde{\mathbf{f}}\|_F + \frac{M_p}{\beta^2} \|\tilde{g}\|_G \quad (328)$$

We proceed by estimating the dual norms $\|\tilde{\mathbf{f}}\|_F$ and $\|\tilde{g}\|_G$. Since for every $\hat{\mathbf{v}} = (\hat{v}_I^r)_{r=1}^n$,

$$\begin{aligned} \hat{\mathbf{v}}^T \tilde{\mathbf{f}} &= \mu \int_{\Omega} \nabla (N_i^u \hat{v}_I) : \nabla (\mathbf{u} - \mathbf{u}_I) \, d\Omega \\ &\quad - \int_{\Omega} \text{div} (N_i^u \hat{v}_I) (p - p_I) \, d\Omega \leq M_u \|\hat{\mathbf{v}}\|_X \|\mathbf{u} - \mathbf{u}_I\|_X \\ &\quad + M_p \|\hat{\mathbf{v}}\|_X \|p - p_I\|_Y \end{aligned} \quad (329)$$

we obtain

$$\frac{\hat{\mathbf{v}}^T \tilde{\mathbf{f}}}{\|\hat{\mathbf{v}}\|_X} \leq M_u \|\mathbf{u} - \mathbf{u}_I\|_X + M_p \|p - p_I\|_Y \quad (330)$$

which gives (cf. the dual norm definition (104))

$$\|\tilde{\mathbf{f}}\|_F \leq M_u \|\mathbf{u} - \mathbf{u}_I\|_X + M_p \|p - p_I\|_Y \quad (331)$$

Analogously, for every $\hat{q} = (\hat{q}_I^r)_{r=1}^m$, we get

$$\hat{q}^T \tilde{g} = - \int_{\Omega} (N_r^p \hat{q}_I) \text{div} (\mathbf{u} - \mathbf{u}_I) \, d\Omega \leq M_p \|\hat{q}\|_Y \|\mathbf{u} - \mathbf{u}_I\|_X \quad (332)$$

and therefore we have

$$\|\tilde{g}\|_G \leq M_p \|\mathbf{u} - \mathbf{u}_I\|_X \quad (333)$$

From (327), (328), (331), and (333) we have

$$\begin{aligned} \|\hat{\mathbf{u}} - \hat{\mathbf{u}}_I\|_X &\leq \left(\frac{M_u}{\alpha} + \frac{M_u^{1/2} M_p}{\alpha^{1/2}\beta} \right) \|\mathbf{u} - \mathbf{u}_I\|_X \\ &\quad + \frac{M_p}{\alpha} \|p - p_I\|_Y \end{aligned} \quad (334)$$

$$\begin{aligned} \|\hat{p} - \hat{p}_I\|_Y &\leq \left(\frac{M_p}{\beta} + \frac{M_p^{3/2}}{\alpha^{1/2}\beta} + \frac{M_u M_p}{\beta^2} \right) \|\mathbf{u} - \mathbf{u}_I\|_X \\ &\quad + M_p \left(\frac{1}{\beta} + \frac{M_p^{1/2}}{\alpha^{1/2}\beta} \right) \|p - p_I\|_Y \end{aligned} \quad (335)$$

Observing that by triangle inequality and Remark 13, it holds

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}^h\|_X &\leq \|\mathbf{u} - \mathbf{u}_I\|_X + \|\mathbf{u}_I - N^u \hat{\mathbf{u}}_I\|_X \\ &= \|\mathbf{u} - \mathbf{u}_I\|_X + \|\hat{\mathbf{u}} - \hat{\mathbf{u}}_I\|_X \end{aligned} \quad (336)$$

and

$$\begin{aligned} \|p - p^h\|_Y &\leq \|p - p_I\|_Y + \|p_I - N^p \hat{p}_I\|_Y \\ &= \|p - p_I\|_Y + \|\hat{p} - \hat{p}_I\|_Y \end{aligned} \quad (337)$$

from (334) and (335), we get the error estimates

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}^h\|_X &\leq \left(1 + \frac{M_u}{\alpha} + \frac{M_u^{1/2} M_p}{\alpha^{1/2}\beta} \right) \|\mathbf{u} - \mathbf{u}_I\|_X \\ &\quad + \frac{M_p}{\alpha} \|p - p_I\|_Y \end{aligned} \quad (338)$$

$$\begin{aligned} \|p - p^h\|_Y &\leq \left(\frac{M_p}{\beta} + \frac{M_p^{3/2}}{\alpha^{1/2}\beta} + \frac{M_u M_p}{\beta^2} \right) \|\mathbf{u} - \mathbf{u}_I\|_X \\ &\quad + \left(1 + \frac{M_p}{\beta} + \frac{M_p^{1/2} M_p}{\alpha^{1/2}\beta} \right) \|p - p_I\|_Y \end{aligned} \quad (339)$$

We notice that the constant M_p , which did not appear in the stability estimates, has now come into play. Furthermore, using (318) and (319), from (338) and (339),

we infer

$$\|u - u^h\|_X + \|p - p^h\|_Y \leq Ch^k \quad (340)$$

with $k = \min(k_u, k_p)$ and $C = C(\alpha, \beta, M_u, M_p)$ independent of h . \square

Remark 18. A crucial step in obtaining error estimate (340) is to prove the bounds (cf. (331) and (333))

$$\|\tilde{u}\|_F \leq M_u \|u - u_I\|_X + M_p \|p - p_I\|_Y \quad (341)$$

$$\|\tilde{p}\|_G \leq M_p \|u - u_I\|_X \quad (342)$$

where \tilde{u} and \tilde{p} are defined by (325) and (326) respectively. The estimates above result from a suitable choice of the norms for $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, $F = \mathbb{R}^n$, and $G = \mathbb{R}^m$. In fact, by choosing for X the norm (269) and for F the corresponding dual norm, we can get (341), as highlighted by (329). Similarly, by choosing for Y the norm (270) and for G the corresponding dual norm, we can obtain (341) (cf. (332)).

Remark 19. The discrete functions u_I and p_I in (317) are typically chosen as follows:

- u_I is the nodal interpolated of u . Therefore,

$$u_I = N_u^T \hat{u}_I \quad (343)$$

where $\hat{u}_I = (\hat{u}_I^e)_{e=1}^E$ is the vector containing the nodal values of u .

- p_I is the projection of p over the pressure approximation space. Therefore,

$$p_I = N_p^T \hat{p}_I \quad (344)$$

where the vector $\hat{p}_I = (\hat{p}_I^e)_{e=1}^E$ is uniquely determined by the following set of m equations

$$\int_{\Omega} N_p^e (N_p^e \hat{p}_I^e) d\Omega = \int_{\Omega} N_p^e p d\Omega \quad s = 1, \dots, m \quad (345)$$

For regular solution (u, p) , standard approximation results (see e.g. Ciarlet, 1978) allow to determine the exponents k_u and k_p entering in estimates (318) and (319) in terms of the selected approximation spaces for the velocity and the pressure fields. For instance, when considering the Crouzeix–Raviart element (cf. Figure 12), we have $k_u = k_p = 2$. Hence, Theorem 5 shows that the discretization error is $O(h^2)$ (see Chapter 4, this Volume).

6 RELATED CHAPTERS

(See also Chapter 4, Chapter 15 of this Volume; Chapter 2, Volume 3).

REFERENCES

- Alotto P and Perugia I. Mixed finite element methods and tree-cotree implicit condensation. *Calcolo* 1999; 36:233–248.
- Amara M and Thomas JM. Equilibrium finite elements for the linear elastic problem. *Numer. Math.* 1979; 33:367–383.
- Arnold DN. Discretization by finite elements of a model parameter dependent problem. *Numer. Math.* 1981; 37:405–421.
- Arnold DN and Brezzi F. Mixed and non-conforming finite element methods: Implementation, post-processing and error estimates. *Math. Modell. Numer. Anal.* 1985; 19:7–35.
- Arnold DN and Winther R. Mixed finite elements for elasticity. *Numer. Math.* 2002; 42:401–419.
- Arnold D, Brezzi F and Douglas J. PEERS: a new mixed finite element for plane elasticity. *Jpn. J. Appl. Math.* 1984; 1:347–367.
- Arnold DN, Brezzi F and Fortin M. A stable finite element for the Stokes equations. *Calcolo* 1984; 21:337–344.
- Arnold DN, Douglas J and Gupta CP. A family of higher order mixed finite element methods for plane elasticity. *Numer. Math.* 1984; 45:1–22.
- Aurini SN, Galligher RH and Zienkiewicz OC. *Hybrid and Mixed Finite Element Methods*. Wiley: New York, 1983.
- Auricchio F, Beirão da Veiga L, Lovadina C and Reali A. Triangular enhanced strain elements for plane linear elasticity. *Comput. Methods Appl. Mech. Eng.*; submitted.
- Baiocchi C and Brezzi F. Stabilization of unstable methods. In *Problemi attuali dell'Analisi e della Fisica Matematica*, Ricci PE (ed.). Università La Sapienza: Roma, 1993; 59–63.
- Baiocchi C, Brezzi F and Franca LP. Virtual bubbles and Ga.L.S. *Comput. Methods Appl. Mech. Eng.* 1993; 105:125–141.
- Baranger J, Maître J-F and Oudin F. Connection between finite volume and mixed finite element methods. *RAIRO Model. Math. Anal. Numér.* 1996; 30:445–465.
- Baùs KJ. *Finite Element Procedures*. Prentice Hall: Englewood Cliffs, NJ, 1996.
- Becker EB, Carey GF and Oden JT. *Finite Elements. An Introduction*. Prentice Hall: Englewood Cliffs, NJ, 1981.
- Behr MA, Franca LP and Tezduyar TE. Stabilized finite element methods for the velocity-pressure-stress formulation of incompressible flows. *Comput. Methods Appl. Mech. Eng.* 1993; 104:31–48.
- Belytschko T, Liu WK and Moran B. *Non Linear Finite Elements for Continua and Structures*. John Wiley & Sons: New York, 2000.
- Bercovier M and Pironneau OA. Error estimates for finite element method solution of the Stokes problem in primitive variables. *Numer. Math.* 1977; 33:211–224.
- Boffi D. Stability of higher order triangular Hood–Taylor methods for the stationary Stokes equations. *Math. Models Methods Appl. Sci.* 1994; 4(2):223–235.
- Boffi D. Minimal stabilizations of the $P_{2,1} - P_1$ approximation of the stationary Stokes equations. *Math. Models Methods Appl. Sci.* 1995; 5(2):213–224.

- Boffi D. Three-dimensional finite element methods for the Stokes problem. *SIAM J. Numer. Anal.* 1997; 34:664–670.
- Boffi D and Lovadina C. Analysis of new augmented Lagrangian formulations for mixed finite element schemes. *Numer. Math.* 1997; 75:405–419.
- Boutet J and Wood RD. *Nonlinear Continuum Mechanics for Finite Element Analysis*. Cambridge University Press: Cambridge, UK, 1997.
- Braess D. *Finite Elements. Theory, Fast Solvers and Applications in Solid Mechanics*. Cambridge University Press, 1997.
- Braess D. Enhanced assumed strain elements and locking in membrane problems. *Comput. Methods Appl. Mech. Eng.* 1998; 165:155–174.
- Brenner SC and Scott LR. *The Mathematical Theory of Finite Element Methods*. Springer-Verlag: New York, 1994.
- Brezzi F. On the existence, uniqueness and approximation of saddle point problems arising from Lagrangian multipliers. *RAIRO Anal. Numer.* 1974; 8:129–151.
- Brezzi F and Douglas J. Stabilized mixed methods for the Stokes problem. *Numer. Math.* 1988; 53:225–235.
- Brezzi F and Falk RS. Stability of higher-order Hood–Taylor Methods. *SIAM J. Numer. Anal.* 1991; 28:581–590.
- Brezzi F and Fortin M. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag: New York, 1991.
- Brezzi F and Fortin M. A minimal stabilisation procedure for mixed finite element methods. *Numer. Math.* 2001; 89:457–492.
- Brezzi F and Pitkäranta J. On the stabilization of finite element approximations of the Stokes equations. In *Efficient Solutions of Elliptic Systems. Notes on Numerical Fluid Mechanics*, vol. 10, Hackbusch W (ed.). Braunschweig Wiesbaden, 1984; 11–19.
- Brezzi F, Douglas J and Marini LD. Two families of mixed finite elements for second order elliptic problems. *Numer. Math.* 1985; 47:217–235.
- Brezzi F, Douglas J and Marini LD. Recent results on mixed finite element methods for second order elliptic problems. In *Vistas in Applied Mathematics. Numerical Analysis, Atmospheric Sciences, Immunology*, Balakrishnan AV, Dorodnitsyn AA and Liou JL (eds). Optimization Software Publications: New York, 1986; 25–43.
- Brezzi F, Douglas J, Fortin M and Marini LD. Efficient rectangular mixed finite elements in two and three space variables. *Math. Modell. Numer. Anal.* 1987; 21:581–604.
- Brezzi F, Douglas J, Duran R and Fortin M. Mixed finite elements for second order elliptic problems in three variables. *Numer. Math.* 1988; 51:237–250.
- Carey GF and Oden JT. *Finite Elements: A Second Course*, vol. II. Prentice Hall: Englewood Cliffs, NJ, 1983.
- Ciarlet PG. *The Finite Element Method for Elliptic Problems. In Stress Analysis. Recent Developments in Numerical and Experimental Methods*, Zienkiewicz OC and Holister GS (eds). John Wiley & Sons: New York, 1965; 85–119.
- Crisfield MA. *Finite Elements and Solution Procedures for Structural Analysis*. Pitman Press: Swansea, UK, 1986.

- Crisfield MA. *Non-Linear Finite Element Analysis of Solids and Structures, Vol. 1 – Essentials*. John Wiley & Sons: New York, 1991.
- Crisfield MA. *Non-Linear Finite Element Analysis of Solids and Structures, Vol. 2 – Advanced Topics*. John Wiley & Sons: New York, 1997.
- Crouzeix M and Raviart PA. Conforming and non-conforming finite element methods for the stationary Stokes equations. *RAIRO Anal. Numer.* 1973; 7:33–76.
- Douglas J and Wang J. An absolutely stabilized finite element method for the Stokes problem. *Math. Comput.* 1989; 52(185):495–508.
- Falk RS. Nonconforming finite element methods for the equations of linear elasticity. *Math. Comput.* 1991; 57:529–550.
- Fortin M. Utilisation de la méthode des éléments finis en mécanique des fluides. *Calcolo* 1975; 12:405–441.
- Fortin M. An analysis of the convergence of mixed finite element methods. *RAIRO Anal. Numer.* 1977; 11:341–354.
- Fraeijs de Veubeke B. Displacement and equilibrium models in the finite element method. In *Stress Analysis. Recent Developments in Numerical and Experimental Methods*, Lectures Notes in Math. 606, Zienkiewicz OC and Holister GS (eds). John Wiley & Sons, 1965; 145–197.
- Fraeijs de Veubeke B. Stress function approach. In *World Congress on the Finite Element Method in Structural Mechanics*, Boummeuth, 1975; 321–352.
- Franca LP and Hughes TJR. Two classes of finite element methods. *Comput. Methods Appl. Mech. Eng.* 1983; 69:89–129.
- Franca LP and Stenberg R. Error analysis of some Galerkin least squares methods for the elasticity equations. *SIAM J. Numer. Anal.* 1991; 28:1680–1697.
- Glowinski R and Pironneau O. Numerical methods for the first biharmonic equation and for the two-dimensional Stokes problem. *SIAM Rev.* 1979; 21:167–212.
- Hansbo P and Larson MG. Discontinuous Galerkin and the Crouzeix–Raviart element: application to elasticity. *Math. Modell. Numer. Anal.* 2003; 37:63–72.
- Hood P and Taylor C. Numerical solution of the Navier–Stokes equations using the finite element technique. *Comput. Fluids* 1973; 1:1–28.
- Hughes TJR. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Prentice Hall: Englewood Cliffs, NJ, 1987.
- Hughes TJR and Franca LP. A new finite element formulation for computational fluid dynamics. VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity/pressure spaces. *Comput. Methods Appl. Mech. Eng.* 1987; 65:85–96.
- Hughes TJR, Franca LP and Balestra M. A new finite element formulation for computational fluid dynamics. V: circumventing the Babuška–Brezzi condition: a stable Petrov–Galerkin formulation of the Stokes problem accommodating equal-order interpolations. *Comput. Methods Appl. Mech. Eng.* 1986; 59:85–99.
- Johnson C. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press: Cambridge, UK, 1992.

- Johnson C and Mercier B. Some equilibrium finite element methods for two-dimensional elasticity problems. *Numer. Math.* 1978; 30:103–116.
- Ladyzhenskaya OA. *The Mathematical Theory of Viscous Incompressible Flow*. Gordon & Breach: New York, 1969.
- Lovadina C. Analysis of strain-pressure finite element methods for the Stokes problem. *Numer. Methods Partial Differ. Equations* 1997; 13:717–730.
- Lovadina C and Auricchio F. On the enhanced strain technique for elasticity problems. *Comput. Struct.* 2003; 18:777–787.
- Mansfield L. Finite element subspaces with optimal rates of convergence for the stationary Stokes problem. *RAIRO Anal. Numer.* 1982; 16:49–66.
- Marini LD. An inexpensive method for the evaluation of the solution of the lowest order Raviart-Thomas mixed method. *SIAM J. Numer. Anal.* 1985; 22:493–496.
- Nedelec JC. Mixed finite elements in \mathbb{R}^3 . *Numer. Math.* 1980; 35:315–341.
- Osseen NS and Peterson H. *Introduction to the Finite Element Method*. Prentice Hall: New York, 1992.
- Panuso D and Bathe KJ. A four-node quadrilateral mixed-interpolated element for solids and fluids. *Math. Models Methods Appl. Sci.* 1995; 5:1113–1128.
- Pierre R. Regularization procedures of mixed finite element approximations of the Stokes problem. *Numer. Methods Partial Differ. Equations* 1989; 5:241–258.
- Quarteroni A and Valli A. *Numerical Approximation of Partial Differential Equations*. Springer-Verlag: New York, 1994.
- Raviart PA and Thomas JM. A mixed finite element method for second order elliptic problems. In *Mathematical Aspects of the Finite Element Method*, Lecture Notes in Mathematics 606, Galligani I and Magenes E (eds). Springer-Verlag: New York, 1977; 292–315.
- Reddy JN. *An Introduction to the Finite Element Method*. McGraw-Hill: New York, 1993.
- Reddy BD and Simo JC. Stability and convergence of a class of enhanced strain methods. *SIAM J. Numer. Anal.* 1995; 32:1705–1728.
- Scott LR and Vogelius M. Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials. *Math. Models Numer. Anal.* 1985; 19:111–143.
- Silvester DJ and Kechar N. Stabilised bilinear-constant velocity-pressure finite elements for the conjugate gradient solution of the Stokes problem. *Comput. Methods Appl. Mech. Eng.* 1990; 79:71–86.
- Simo JC. Topics on the numerical analysis and simulation of plasticity. In *Handbook of numerical analysis*, vol. III, Ciarlet PG and Lions JL (eds). Elsevier Science Publisher B.V., 1999; 193–499.
- Simo JC and Hughes TJR. *Computational Inelasticity*. Springer-Verlag: New York, 1998.
- Simo JC and Rifai MS. A class of mixed assumed strain methods and the method of incompatible modes. *Int. J. Numer. Methods Eng.* 1990; 29:1595–1638.
- Stenberg R. Analysis of mixed finite element methods for the Stokes problem: a unified approach. *Math. Comput.* 1984; 42:9–23.
- Stenberg R. On some three-dimensional finite elements for incompressible media. *Comput. Methods Appl. Mech. Eng.* 1987; 63:261–269.
- Stenberg R. A family of mixed finite elements for the elasticity problem. *Numer. Math.* 1988; 53:513–538.
- Strang G and Fix GJ. *An Analysis of the Finite Element Method*. Prentice Hall: Englewood Cliffs, NJ, 1973.
- Temam R. *Navier-Stokes Equations*. North Holland: Amsterdam, 1977.
- Turner MJ, Clough RW, Martin HC and Topp LJ. Stiffness and deflection analysis of complex structures. *J. Aeronaut. Sci.* 1956; 23:805–823.
- Verfürth R. Error estimates for a mixed finite element approximation of the Stokes equation. *RAIRO Anal. Numer.* 1984; 18:175–182.
- Vogelius M. A right-inverse for the divergence operator in spaces of piecewise polynomials. *Numer. Math.* 1983; 41:19–37.
- Wait R and Mitchell AR. *Finite Element Analysis and Applications*. John Wiley & Sons: Chichester, West Sussex, 1985.
- Zienkiewicz OC and Taylor RL. *The Finite Element Method* (5th edn), Vol. 1 – *The Basis*. Butterworth-Heinemann: Oxford, 2000a.
- Zienkiewicz OC and Taylor RL. *The Finite Element Method*, (5th edn), Vol. 2 – *Solid Mechanics*. Butterworth-Heinemann: Oxford, 2000b.
- Zienkiewicz OC and Taylor RL. *The Finite Element Method* (5th edn), Vol. 3 – *Fluid Dynamics*. Butterworth-Heinemann: Oxford, 2000c.
- Zienkiewicz OC, Taylor RL and Baynham JAW. Mixed and irreducible formulations in finite element analysis. In *Hybrid and Mixed Finite Element Methods*, Atluri SN, Gallagher RH and Zienkiewicz OC (eds). John Wiley & Sons: New York, 1983; 405–431, Chap. 21.

Chapter 10

Meshfree Methods

Antonio Huerta², Ted Belytschko¹, Sonia Fernández-Méndez² and Timon Rabczuk¹

¹Northwestern University, Evanston, IL, USA

²Universitat Politècnica de Catalunya, Barcelona, Spain

| | |
|--|-----|
| 1 Introduction | 279 |
| 2 Approximation in Meshfree Methods | 280 |
| 3 Discretization of Partial Differential Equations | 291 |
| 4 Radial Basis Functions | 300 |
| 5 Discontinuities | 300 |
| 6 Blending Meshfree Methods and Finite Elements | 303 |
| References | 306 |

1 INTRODUCTION

As the range of phenomena that need to be simulated in engineering practice broadens, the limitations of conventional computational methods, such as finite elements, finite volumes, or finite difference methods, have become apparent. There are many problems of industrial and academic interest that cannot be easily treated with these classical mesh-based methods: for example, the simulation of manufacturing processes such as extrusion and molding, where it is necessary to deal with extremely large deformations of the mesh, or simulations of failure, where the simulation of the propagation of cracks with arbitrary and complex paths is needed.

The underlying structure of the classical mesh-based methods is not well suited to the treatment of discontinuities

that do not coincide with the original mesh edges. With a mesh-based method, a common strategy for dealing with moving discontinuities is to remesh whenever it is necessary. The remeshing process is costly and not trivial in 3D (if reasonable meshes are desired), and projection of quantities of interest between successive meshes usually leads to degradation of accuracy and often results in an excessive computational cost. Although some recent developments (Moes, Dolbow and Belytschko, 1999; Wells, Borst and Sluys, 2002) partially overcome these difficulties, the implementation of discontinuities is not as simple as in meshfree methods.

The objective of meshfree methods is to eliminate at least part of this mesh dependence by constructing the approximation entirely in terms of nodes (usually called particles in the context of meshfree methods). Moving discontinuities or interfaces can usually be treated without remeshing with minor costs and accuracy degradation (see, for instance, Belytschko and Organ, 1997). Thus the range of problems that can be addressed by meshfree methods is much wider than mesh-based methods. Moreover, large deformations can be handled more robustly with meshfree methods because the approximation is not based on elements whose distortion may degrade the accuracy. This is useful in both fluid and solid computations.

Another major drawback of mesh-based methods is the difficulty in ensuring for any real geometry, a smooth, painless, and seamless integration with computer aided engineering (CAE), industrial computer aided design (CAD), and computer aided manufacturing (CAM) tools. Meshfree methods have the potential to circumvent these difficulties.

The elimination of mesh generation is the key issue. The advantages of meshfree methods for 3D computations become particularly apparent.

Meshfree methods also present obvious advantages in adaptive processes. There are a priori error estimates for most of the meshfree methods. This allows the definition of adaptive refinement processes as in finite element computations: an *a posteriori* error estimate is computed and the solution is improved by adding nodes/particles where needed or increasing the order of the approximation until the error becomes acceptable (see e.g. Babuska and Melenk, 1995; Melenk and Babuska, 1996; Duarte and Oden, 1996a; Babuska, Banerjee and Osborn, 2002a).

Meshfree methods were originated over 25 years ago but it is in recent years that they have received substantial attention. The approach that seems to have the longest continuous history is the SPH method by Lucy (1977) and Gingold and Monaghan (1977) (see Section 2.1). It was first developed for modelling astrophysical phenomena without boundaries, such as exploding stars and dust clouds. Compared to other numerical methods, the rate of publications in this field was very modest for many years; progress is reflected in the review papers by Monaghan (1982, 1988).

Recently, there has been substantial improvement in these methods. For instance, Dyka (1994) and Sweegle, Hicks and Attaway (1995) study its instabilities, Johnson and Beissel (1996) propose a method for improving strain calculations, and Liu, Jun and Zhang (1995b) present a correction function for kernels in both the discrete and continuous case.

In fact, this approach can be seen as a variant of MLS approximations (see Section 2.2). A detailed description of MLS approximations can be found in Lancaster (1981). Nayroles, Touzot and Villon (1992) were evidently the first to use moving least square approximations in a Galerkin weak form and called it the *diffuse element method* (DEM). Belytschko, Lu and Gu (1994) refined the method and extended it to discontinuous approximations and called it *element-free Galerkin* (EFG). Duarte and Oden (1996b) and Babuska and Melenk (1995) recognize that the methods are specific instances of *partitions of unity* (PU), a method first proposed in Babuska, Caloz and Osborn (1994). Duarte and Oden (1996a) and Liu, Li and Belytschko (1997a) were also among the first to prove convergence of meshfree methods.

This class of methods (EFG, DEM, PU, among others) is consistent and in the forms proposed stable, although substantially more expensive than SPH because of the need of a very accurate integration. Zhu and Atluri (1998) propose a Petrov-Galerkin weak form in order to facilitate the computation of the integrals, but usually leading to nonsymmetric systems of equations. De and Bathe (2000) use this

approach for a particular choice of the approximation space and the Petrov-Galerkin weak form and call it the *method of finite spheres*.

On a parallel path, Vila (1999) has introduced a different meshfree approximation specially suited for conservation laws: the *renormalized meshless derivative* (RMD) which turns out to give accurate approximation of derivatives in the framework of collocation approaches. Two other paths in the evolution of meshfree methods have been the development of generalized finite difference methods, which can deal with arbitrary arrangements of nodes, and particle-in-cell methods. One of the early contributors to the former was Perrone and Kao (1975), but Liszka and Orkisz (1980) proposed a more robust method. Recently, these methods have taken a character that closely resembles the moving least squares methods.

In recent papers, the possibilities of meshfree methods have become apparent. The special issue Liu, Belytschko, and Oden (1996a) shows the ability of meshfree methods to handle complex simulations, such as impact, cracking, or fluid dynamics. Bouillard and Suleau (1998) apply a meshfree formulation to acoustic problems with good results. Bonet and Lok (1999) introduce a gradient correction in order to preserve the linear and angular momentum with applications to fluid dynamics. Bonet and Kulasegaram (2000) proposes the introduction of integration correction that improves accuracy with applications to metal forming simulation. Oñate and Idelsohn (1998) propose a meshfree method, the *finite point method*, based on a weighted least-squares approximation with point collocation with applications to convective transport and fluid flow. Recently several authors have proposed mixed approximations combining finite elements and meshfree methods, in order to exploit the advantages of each method (see Belytschko, Organ and Krongauz, 1995; Hegen, 1996; Liu, Uras and Chen, 1997b; Huerta and Fernández-Méndez, 2000a; Hao, Liu and Belytschko, 2004). Several review papers and books have been published on meshfree methods (see Belytschko *et al.*, 1996b; Li and Liu, 2002; Babuska, Banerjee and Osborn, 2003; Liu *et al.*, 1995a). Two recent books are Atluri and Shen (2002) and Liu (2002).

2 APPROXIMATION IN MESHFREE METHODS

This section describes the most common approximants in meshfree methods. We will employ the name 'approximants' rather than interpolants that is often mistakenly used in the meshfree literature because, as shown later, these approximants usually do not pass through the data, so they are not interpolants. Meshfree approximants can

be classified into two families: those based on SPH and those based on MLS. As will be noted in Section 3, the SPH approximants are usually combined with collocation or point integration techniques, while the MLS approximants are customarily applied with Galerkin formulations, though collocation techniques are growing in popularity.

2.1 Smooth particle hydrodynamic

2.1.1 The early SPH

The earliest meshfree method is the SPH method (see Lucy, 1977). The basic idea is to approximate a function $u(x)$ by a convolution

$$u(x) \approx \bar{u}^\rho(x) := \int C_\rho \phi\left(\frac{y-x}{\rho}\right) u(y) dy \quad (1)$$

where ϕ is a compactly supported function, usually called a *window function* or *weight function*, and ρ is the so-called dilation parameter. The support of the function is sometimes called the domain of influence. The dilation parameter characterizes the size of the support of $\phi(x/\rho)$, usually by its radius. C_ρ is a normalization constant such that

$$\int C_\rho \phi\left(\frac{y}{\rho}\right) dy = 1 \quad (2)$$

One way to develop a discrete approximation from (1) is to use numerical quadrature

$$u(x) \approx \bar{u}^\rho(x) \approx u^\rho(x) := \sum_i C_\rho \phi\left(\frac{x_i-x}{\rho}\right) u(x_i) \omega_i$$

where x_i and ω_i are the points and weights of the numerical quadrature. The quadrature points are usually called *particles*. The previous equation can also be written as

$$u(x) \approx \bar{u}^\rho(x) \approx u^\rho(x) := \sum_i \omega(x_i, x) u(x_i) \quad (3)$$

where the discrete window function is defined as

$$\omega(x_i, x) = C_\rho \phi\left(\frac{x_i-x}{\rho}\right) \omega_i$$

Thus, the SPH meshfree approximation can be defined as

$$u(x) \approx u^\rho(x) = \sum_i N_i(x) u(x_i)$$

with the approximation basis $N_i(x) = \omega(x_i, x)$.

Remark 1. Note that, in general, $u^\rho(x_i) \neq u(x_i)$. That is, the shape functions are not interpolants, that is, they do not verify the Kronecker delta property:

$$N_j(x_i) \neq \delta_{ij}$$

This is common for all particle methods (see Figure 5 for the MLS approximant) and thus special techniques are needed to impose essential boundary conditions (see Section 3).

Remark 2. The dilation parameter ρ characterizes the support of the approximants $N_i(x)$.

Remark 3. In contrast to finite elements, the neighbor particles (particles belonging to a given support) have to be identified during the course of the computation. This is of special importance if the domain of support changes in time and requires fast neighbor search algorithms, a crucial feature for the effectiveness of a meshfree method (see e.g. Schweizer, 2003).

Remark 4. There is an optimal value for the ratio between the dilation parameter ρ and the distance between particles h . Figure 1 shows that for a fixed distribution of particles, h constant, the dilation parameter must be large enough to avoid aliasing (spurious short waves in the approximated solution). It also shows that an excessively large value for ρ will lead to excessive smoothing. For this reason, it is usual to maintain a constant ratio between the dilation parameter ρ and the distance between particles h .

2.1.2 Window functions

The window function plays an important role in meshfree methods. Other names for the window function are kernel and weight function. The window function may be defined in various manners. For 1D the most common choices are *Cubic spline*:

$$\phi_{1D}(x) = 2 \begin{cases} 4(|x|-1)x^2 + (2/3) & |x| \leq 0.5 \\ 4(1-|x|)^3/3 & 0.5 \leq |x| \leq 1 \\ 0 & 1 \leq |x| \end{cases} \quad (4)$$

Gaussian:

$$\phi_{1D}(x) = \begin{cases} \frac{\exp(-9x^2) - \exp(-9)}{1 - \exp(-9)} & |x| \leq 1 \\ 0 & 1 \leq |x| \end{cases} \quad (5)$$

The above window function can easily be extended to higher dimensions. For example, in 2D the most common extensions are

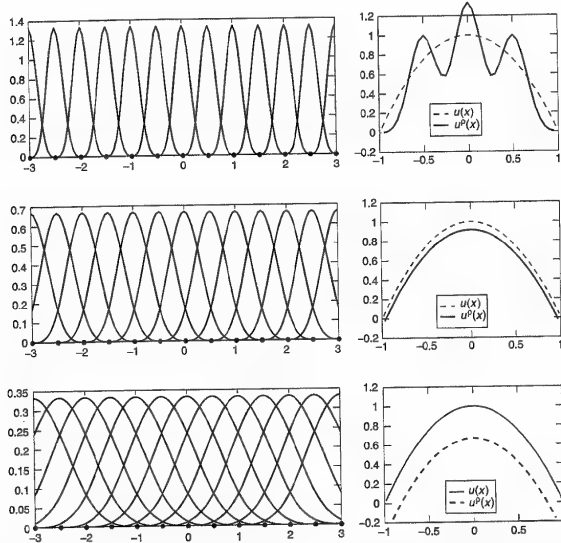


Figure 1. SPH approximation functions and approximation of $u(x) = 1 - x^2$ with cubic spline window function, distance between particles $h = 0.5$, and quadrature weights $w_i = h$, for $\rho/h = 1, 2, 4$.

Window function with spherical (cylindrical) support:

$$\phi(x) = \phi_{1D}(\|x\|)$$

Window function with rectangular support (tensor product):

$$\phi(x) = \phi_{1D}(|x_1|) \phi_{1D}(|x_2|)$$

where, as usual, $x = (x_1, x_2)$ and $\|x\| = \sqrt{x_1^2 + x_2^2}$.

2.1.3 Design of window functions

In the continuous SPH approximation (1), a window function ϕ can easily be modified to exactly reproduce a

polynomial space \mathcal{P}_m in \mathbb{R} of degree $\leq m$, that is,

$$p(x) = \int C_p \phi\left(\frac{y-x}{\rho}\right) p(y) dy, \quad \forall p \in \mathcal{P}_m \quad (6)$$

If the following conditions are satisfied

$$\int C_p \phi\left(\frac{y}{\rho}\right) dy = 1,$$

$$\int C_p \phi\left(\frac{y}{\rho}\right) y^j dy = 0 \quad \text{for } 0 < j \leq m$$

the window function is able to reproduce the polynomial space \mathcal{P}_m . Note that the first condition coincides with (2) and defines the normalization constant, that is, it imposes

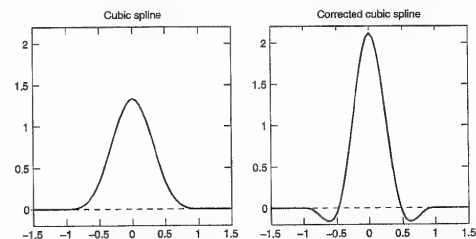


Figure 2. Cubic spline and corrected window function for polynomials of degree 2.

the reproducibility of constant functions. The ability to reproduce functions of order m is often referred as m th order consistency.

For example, the window function

$$\bar{\phi}(x) = \left(\frac{27}{17} - \frac{120}{17}x^2\right)\phi(x) \quad (7)$$

where $\phi(x)$ is the cubic spline, reproduces the second degree polynomial basis $\{1, x, x^2\}$. Figure 2 shows the cubic spline defined in (4) and the corrected window function (7) (see Liu *et al.*, 1996b for details).

However, the design of the window function is not trivial in the presence of boundaries or with nonuniform distributions of particles (see Section 2.2). Figure 3 shows the corrected cubic spline window functions associated with a uniform distribution of particles, with distance $h = 0.5$ between particles and $\rho = 2h$, and the discrete SPH approximation described by (3) for $u(x) = x$ with uniform weights $w_i = h$. The particles outside the interval $[-1, 1]$ are considered in the approximation, as in an unbounded

domain (the corresponding translated window functions are depicted with a dashed line). The linear monomial $u(x) = x$ is exactly reproduced. However, in a bounded domain, the approximation is not exact near the boundaries when only the particles in the domain $[-1, 1]$ are considered in this example (see Figure 4).

Remark 5 (Consistency) If the approximation reproduces exactly a basis of the polynomials of degree less or equal to m then the approximation is said to have m -order consistency.

2.1.4 Correcting the SPH method

The SPH approximation is used in the solution of PDEs, usually through a collocation technique or point integration approaches (see Monaghan, 1982; Vila, 1999; Bonet and Lok, 1999; and Section 3.1). Thus, it is necessary to compute accurate approximations of the derivatives of the dependent variables. The derivatives provided by original

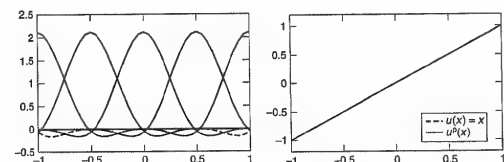


Figure 3. Modified cubic splines and particles, $h = 0.5$, and SPH discrete approximation for $u(x) = x$ with $\rho/h = 2$ in an 'unbounded domain'.

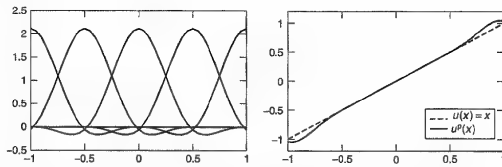


Figure 4. Modified cubic splines and particles, $h = 0.5$, and SPH discrete approximation for $u(x) = x$ with $\rho/h = 2$ in a bounded domain.

SPH method can be quite inaccurate, and thus, it is necessary to improve the approximation, or its derivatives, in some manner.

Randles and Libersky (1996), Krongauz and Belytschko (1998a), and Vila (1999) proposed a correction of the gradient: the RMD. It is an extension of the partial correction proposed by Johnson and Beissel (1996). Let the derivatives of a function u be approximated as the derivatives of the SPH approximation defined in (3),

$$\nabla u(x) \approx \nabla u^p(x) = \sum_j \nabla \omega(x_j, x) u(x_j)$$

The basic idea of the RMD approximation is to define a corrected derivative

$$D^p u(x) := \sum_j B(x) \nabla \omega(x_j, x) u(x_j) \quad (8)$$

where the correction matrix $B(x)$ is chosen such that $\nabla u(x) = D^p u(x)$ for all linear polynomials. In Vila (1999), a symmetrized approximation for the derivatives is defined

$$D_s^p u(x) := D^p u(x) - D^p 1(x) u(x) \quad (9)$$

where, by definition (8),

$$D^p 1(x) = \sum_j B(x) \nabla \omega(x_j, x)$$

Note that (9) exactly interpolates the derivatives when $u(x)$ is constant. The consistency condition $\nabla u(x) = D_s^p u(x)$ must be imposed only for linear monomials

$$B(x) = \left[\sum_j \nabla \omega(x_j, x) (x_j - x)^T \right]^{-1}$$

If the ratio between the dilation parameter ρ and the distance between particles remains constant, there are a

priori error bounds for the RMD, $D_s^p u$, similar to the linear finite element ones, where ρ plays the role of the element size in finite elements (see Vila, 1999).

In SPH, the interparticle forces coincide with the vector joining them, so conservation of linear and angular momentum are met for each point pair. In other words, linear and angular momentum are conserved locally (see Dilts, 1999).

When the kernels are corrected to reproduce linear polynomials or the derivatives of linear polynomials, these local conservation properties are lost. However, global linear and translational momentum are conserved if the approximation reproduces linear functions (see Krongauz and Belytschko, 1998a and Bonet and Lok, 1999).

Although the capability to reproduce a polynomial of a certain order is an ingredient in many convergence proofs of solutions for PDEs, it does not always suffice to pass the patch test. Krongauz and Belytschko (1998b) found that corrected gradient methods do not satisfy the patch test and exhibit poor convergence when the corrected gradient is used for the test function. They showed that a Petrov-Galerkin method with Shepard test functions satisfies the patch test.

There are other ways of correcting the SPH method. For example Bonet and Lok (1999) combine a correction of the window function, as in the reproducing kernel particle method (RKPM) method (see Section 2.2), and a correction of the gradient to preserve angular momentum. In fact, there are a lot of similarities between the corrected SPH and the renormalized meshless derivative. The most important difference between the RMD approach, where the 0-order consistency is obtained by the definition of the symmetrized gradient (8), and the corrected gradient $\tilde{\nabla} u^p$ is that in this case 0-order consistency is obtained with the Shepard function.

With a similar rationale, Bonet and Kulasegaram (2000) present a correction of the window function and an integration corrected vector for the gradients (in the context of metal forming simulations). The corrected approximation

is used in a weak form with numerical integration at the particles (see Section 3.2). Thus, the gradient must be evaluated only at the particles. However, usually the particle integration is not accurate enough and the approximation fails to pass the patch test. In order to obtain a consistent approximation, a corrected gradient is defined. At every particle x_j , the corrected gradient is computed as

$$\tilde{\nabla} u^p(x_j) = \nabla u^p(x_j) + \gamma_j \llbracket u \rrbracket_j$$

where γ_j is the correction vector (one component for each spatial dimension) at particle x_j and where the bracket $\llbracket u \rrbracket_j$ is defined as $\llbracket u \rrbracket_j = u(x_j) - u^p(x_j)$. These extra parameters, γ_j , are determined requiring that the patch test be passed. A global linear system of equations must be solved to compute the correction vector and to define the derivatives of the approximation; then, the approximation of u and its derivatives are used to solve the boundary value problem.

2.2 Moving least-squares approximants

2.2.1 Continuous moving least-squares

The objective of the MLS approach is to obtain an approximation similar to a SPH approximant (1), with high accuracy even in a bounded domain. Let us consider a bounded or unbounded domain Ω . The basic idea of the MLS approach is to approximate $u(x)$, at a given point x , through a polynomial least-squares fit of u in a neighborhood of x . That is, for fixed $x \in \Omega$, and z near x , $u(z)$ is approximated with a polynomial expression

$$u(z) \approx \tilde{u}^p(z, x) = P^T(z) c(x) \quad (10)$$

where the coefficients $c(x) = \{c_0(x), c_1(x), \dots, c_l(x)\}^T$ are not constant, they depend on point x , and $P(z) = \{p_0(z), p_1(z), \dots, p_l(z)\}^T$ includes a complete basis of the subspace of polynomials of degree m . It can also include exact features of a solution, such as cracktip fields, as described in Fleming *et al.* (1997). The vector $c(x)$ is obtained by a least-squares fit, with the scalar product

$$\langle f, g \rangle_x = \int_{\Omega} \phi\left(\frac{y-x}{\rho}\right) f(y) g(y) dy \quad (11)$$

That is, the coefficients c are obtained by minimization of the functional $\tilde{J}_x(c)$ centered in x and defined by

$$\tilde{J}_x(c) = \int_{\Omega} \phi\left(\frac{y-x}{\rho}\right) [u(y) - P(y) c(x)]^2 dy \quad (12)$$

where $\phi((y-x)/\rho)$ is the compact supported weighting function. The same weighting/window functions as for SPH, given in Section 2.1.2, are used.

Remark 6. Thus, the scalar product is centered at the point x and scaled with the dilation parameter ρ . In fact, the integration is constructed in a neighborhood of radius ρ centered at x , that is, in the support of $\phi((y-x)/\rho)$.

Remark 7 (Polynomial space) In one dimension, we can let $p_l(x)$ be the monomials x^l and, in this particular case, $l = m$. For larger spatial dimensions, two types of polynomial spaces are usually chosen: the set of polynomials \mathcal{P}_m of total degree $\leq m$, and the set of polynomials \mathcal{Q}_m of degree $\leq m$ in each variable. Both include a complete basis of the subspace of polynomials of degree m . This, in fact, characterizes the a priori convergence rate (see Liu, Li and Belytschko, 1997a or Fernández-Méndez, Díez and Huerta, 2003).

The vector $c(x)$ is the solution of the normal equations, that is, the linear system of equations

$$M(x) c(x) = \{P, u\}_x \quad (13)$$

where $M(x)$ is the Gram matrix (sometimes called a moment matrix),

$$M(x) = \int_{\Omega} \phi\left(\frac{y-x}{\rho}\right) P(y) P^T(y) dy \quad (14)$$

From (14) and (10), the least-squares approximation of u in a neighborhood of x is

$$u(z) \approx \tilde{u}^p(z, x) = P^T(z) M^{-1}(x) \{P, u\}_x \quad (15)$$

Since the weighting function ϕ usually favors the central point x , it seems reasonable to assume that such an approximation is more accurate precisely at $z = x$ and thus the approximation (15) is particularized at x , that is,

$$u(x) \approx \tilde{u}^p(x) := \tilde{u}^p(x, x)$$

with

$$\tilde{u}^p(x, x) = \int_{\Omega} \phi\left(\frac{y-x}{\rho}\right) P^T(x) M^{-1}(x) P(y) u(y) dy \quad (16)$$

where the definition of the scalar product, equation (11), has been explicitly used. Equation (16) can be rewritten as

$$u(x) \approx \tilde{u}^p(x) = \int_{\Omega} C_p(y, x) \phi\left(\frac{y-x}{\rho}\right) u(y) dy$$

which is similar to the SPH approximation (see equation (1)) and with the scalar correction term $C_p(y, x)$ defined as

$$C_p(y, x) := P^T(x) M^{-1}(x) P(y)$$

The function defined by the product of the correction and the window function ϕ ,

$$\Phi(y, x) := C_\rho(y, x) \phi\left(\frac{y-x}{\rho}\right)$$

is usually called *kernel function*. The new correction term depends on the point x and the integration variable y ; it provides an accurate approximation even in the presence of boundaries (see Liu *et al.*, 1996b for more details). In fact, the approximation verifies the following consistency property (see also Wendland, 2001).

Proposition 1 (Consistency/reproducibility property) The MLS approximation exactly reproduces all the polynomials in \mathbf{P} .

Proof. The MLS approximation of the polynomials $p_i(x)$ is

$$\tilde{p}_i^l(x) = \mathbf{P}^T(x) \mathbf{M}^{-1}(x) (\mathbf{P}, p_i)_x \quad \text{for } i = 0, \dots, l$$

or, equivalently, in vector form

$$[\tilde{\mathbf{P}}^l]^T(x) = \mathbf{P}^T(x) \mathbf{M}^{-1}(x) \int_{\Omega} \phi\left(\frac{y-x}{\rho}\right) \mathbf{P}(y) \mathbf{P}^T(y) dy$$

$\mathbf{M}(x)$

Therefore, using the definition of \mathbf{M} (see (14)), it is trivial to verify that $\tilde{\mathbf{P}}^l(x) = \mathbf{P}(x)$. \square

2.2.2 Reproducing kernel particle method approximation

Application of a numerical quadrature in (16) leads to the RKPM approximation

$$u(x) \simeq \tilde{u}^p(x) \simeq u^p(x) := \sum_{I \in S_x^p} \omega(x_I, x) \mathbf{P}^T(x) \mathbf{M}^{-1}(x) \mathbf{P}(x_I) u(x_I)$$

where $\omega(x_I, x) = \omega_I \phi((x_I - x)/\rho)$ and x_I and ω_I are integration points (particles) and weights respectively. The particles cover the computational domain Ω , $\Omega \subset \mathbb{R}^m$. Let S_x^p be the index set of particles whose support include the point x (see Remark 9). This approximation can be written as

$$u(x) \simeq u^p(x) = \sum_{I \in S_x^p} N_I(x) u(x_I) \quad (17)$$

where the approximation functions are defined by

$$N_I(x) = \omega(x_I, x) \mathbf{P}^T(x) \mathbf{M}^{-1}(x) \mathbf{P}(x_I) \quad (18)$$

Remark 8. In order to preserve the consistency/reproducibility property, the matrix \mathbf{M} , defined in (14), must be evaluated with the same quadrature used in the discretization of (16) (see Chen *et al.*, 1996 for details). That is, matrix $\mathbf{M}(x)$ must be computed as

$$\mathbf{M}(x) = \sum_{I \in S_x^p} \omega(x_I, x) \mathbf{P}(x_I) \mathbf{P}^T(x_I) \quad (19)$$

Remark 9. The sums in (17) and (19) only involve the indices I such that $\phi(x_I, x) \neq 0$, that is, particles x_I in a neighborhood of x . Thus, the set of neighboring particles is defined by the set of indices

$$S_x^p := \{I \text{ such that } \|x_I - x\| \leq \rho\}$$

Remark 10 (Conditions on particle distribution) The matrix $\mathbf{M}(x)$ in (19) must be regular at every point x in the domain. Liu, Li and Belytschko (1997a) discuss the necessary conditions. In fact, this matrix can be viewed (see Huerta and Fernández-Méndez, 2000a or Fernández-Méndez and Huerta, 2002), as a Gram matrix defined with a discrete scalar product

$$(f, g)_x = \sum_{I \in S_x^p} \omega(x_I, x) f(x_I) g(x_I) \quad (20)$$

If this scalar product is degenerated, $\mathbf{M}(x)$ is singular. Regularity of $\mathbf{M}(x)$ is ensured by having enough particles in the neighborhood of every point x and avoiding degenerate patterns, that is

- (i) $\text{card } S_x^p \geq l + 1$.
- (ii) $\mathcal{B}F \in \text{span}\{p_0, p_1, \dots, p_l\} \setminus \{0\}$ such that $F(x_I) = 0 \forall I \in S_x^p$.

Condition (ii) is easily verified. For instance, for $m = 1$ (linear interpolation), the particles cannot lie in the same straight line or plane for, respectively, 2D and 3D. In 1D, for any value of m , it suffices that different particles do not have the same position. Under these conditions, one can compute the vector $\mathbf{P}^T(x) \mathbf{M}^{-1}(x)$ at each point and thus determine, from (18), the shape functions, $N_I(x)$.

2.2.3 Discrete MLS: element-free Galerkin approximation

The MLS development, already presented in Section 2.2.1 for the continuous case (i.e. using integrals), can be developed directly from a discrete formulation. As in the continuous case, the idea is to approximate $u(x)$, at a given point x , by a polynomial least-squares fit of u in a neighborhood of x . That is, the same expression presented in (10) can

be used, namely, for fixed $x \in \Omega$, and z near x ; $u(z)$ is approximated with the polynomial expression

$$u(z) \simeq u^p(z, x) = \mathbf{P}^T(z) \mathbf{c}(x) \quad (21)$$

In the framework of the EFG method, the vector $\mathbf{c}(x)$ is also obtained by a least-squares fit, with the discrete scalar product defined in (20), where $\omega(x_I, x)$ is the discrete weighting function, which is equivalent to the window function

$$\omega(x_I, x) = \phi\left(\frac{x_I - x}{\rho}\right) \quad (22)$$

and S_x^p is the set of indices of neighboring particles defined in Remark 9. That is, the coefficients \mathbf{c} are obtained by minimization of the discrete functional $J_x(\mathbf{c})$ centered in x and defined by

$$J_x(\mathbf{c}) = \sum_{I \in S_x^p} \omega(x_I, x) [u(x_I) - \mathbf{P}(x_I) \mathbf{c}(x)]^2 \quad (23)$$

The normal equations are defined in a similar manner,

$$\mathbf{M}(x) \mathbf{c}(x) = (\mathbf{P}, u)_x \quad (24)$$

and the Gram matrix is directly obtained from the discrete scalar product (see equation (19)). After substitution of the solution of (24) in (21), the least-squares approximation of u in a neighborhood of x is obtained

$$u(z) \simeq u^p(z, x) = \mathbf{P}^T(z) \mathbf{M}^{-1}(x) \sum_{I \in S_x^p} \omega(x_I, x) \mathbf{P}(x_I) u(x_I) \quad (25)$$

Particularization of (25) at $z = x$ leads to the discrete MLS approximation of $u(x)$

$$u(x) \simeq u^p(x) := u^p(x, x) \text{ with } u^p(x, x) = \sum_{I \in S_x^p} \omega(x_I, x) \mathbf{P}^T(x) \mathbf{M}^{-1}(x) \mathbf{P}(x_I) u(x_I) \quad (26)$$

This EFG approximation coincides with the RKPM approximation described in equations (18) and (19).

Remark 11 (Convergence) Liu, Li and Belytschko (1997a) showed convergence of the RKPM and EFG. The a priori error bound is very similar to the bound in finite elements. The parameter ρ plays the role of h and m (the order of consistency) plays the role of the degree of the approximation polynomials in the finite element mesh. Convergence properties depend on m and ρ . They depend on the distance between particles because usually this distance is proportional to ρ , that is, the ratio between the particle distance over the dilation parameter is of order one (see Liu, Li and Belytschko, 1997a).

Remark 12. The approximation is characterized by the order of consistency required, that is, the complete basis of polynomials employed in \mathbf{P} , and by the ratio between the dilation parameter and the particle distance, ρ/h . In fact, the bandwidth of the stiffness matrix increases with the ratio ρ/h (more particles lie inside the circle of radius ρ) (see for instance Figure 5). Note that, for linear consistency, when ρ/h goes to 1, the linear finite element shape functions are recovered.

Remark 13 (Continuity) If the weight function ϕ is C^k , then the EFG/MLS shape functions and the RKPM shape functions are C^k (see Liu, Li and Belytschko, 1997a). Thus, if the window function is a cubic spline, as shown in Figures 6 and 7, the first and second derivatives of the shape functions are well defined throughout the domain, even with linear consistency.

2.2.4 Reproducibility of the MLS approximation

The MLS shape functions can be also obtained by imposing a priori the reproducibility properties of the approximation. Consider a set of particles x_I and a complete polynomial base $\mathbf{P}(x)$. Let us assume an approximation of the form

$$u(x) \simeq \sum_{I \in S_x^p} N_I(x) u(x_I) \quad (27)$$

with approximation functions defined as

$$N_I(x) = \omega(x_I, x) \mathbf{P}^T(x_I) \alpha(x) \quad (28)$$



Figure 5. Interpolation functions with $\rho/h \simeq 1$ (similar to finite elements) and $\rho/h = 2.6$, with cubic spline and linear consistency.

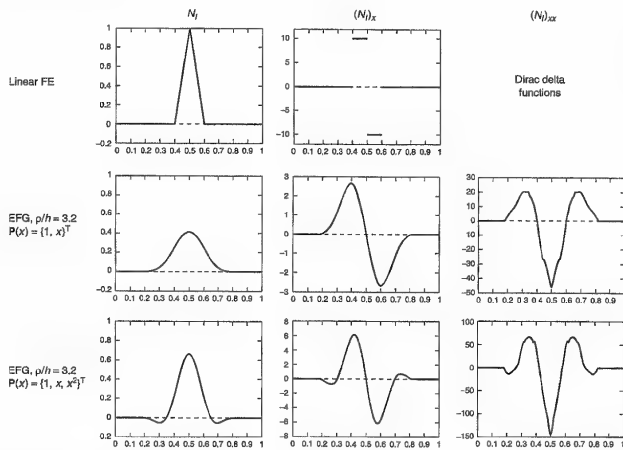


Figure 6. Shape function and derivatives for linear finite elements and the EFG approximation. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecom>

The unknown vector $\alpha(x)$ in \mathbb{R}^{t+1} is determined by imposing the reproducibility condition, which imposes that the approximation proposed in (27) is exact for all the polynomials in P , namely,

$$P(x) = \sum_{j \in S_x^t} N_j(x) P(x_j) \quad (29)$$

After substitution of (28) in (29), the linear system of equations that determines $\alpha(x)$ is obtained

$$M(x) \alpha(x) = P(x) \quad (30)$$

That is,

$$\alpha(x) = M^{-1}(x) P(x) \quad (31)$$

where $M(x)$ is the same matrix defined in (19). Finally, the approximation functions $N_j(x)$ are defined by substituting in (28) the vector α (see (31)). Note that, with this substitution, the expression (18) for the MLS approximation

functions is recovered and consistency is ensured by construction.

Section 2.2.7 is devoted to some implementation details of the EFG method. In particular, it is shown that the derivatives of $N_j(x)$ can be computed without excessive overhead.

2.2.5 MLS centered and scaled approach

For computational purposes, it is usual and preferable to center in x_j and scale with ρ the polynomials involved in the definition of the meshfree approximation functions (see Liu, Li and Belytschko, 1997a or Huerta and Fernández-Méndez, 2000a). Thus, another expression for the EFG shape functions is employed

$$N_j(x) = \omega(x_j, x) P\left(\frac{x_j - x}{\rho}\right) \alpha(x) \quad (32)$$

which is similar to (28). Recall also that typical expressions for the window function are of the type $\phi(y, x) = \phi((y -$

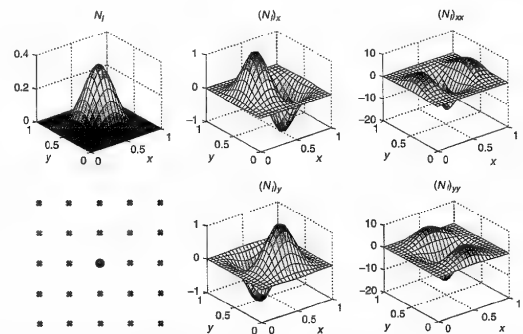


Figure 7. Distribution of particles, EFG approximation function and derivatives, with $\rho/h \approx 2.2$ with circular supported cubic spline and linear consistency. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecom>

$x)/\rho$). The consistency condition becomes in this case

$$P(0) = \sum_{j \in S_x^t} N_j(x) P\left(\frac{x_j - x}{\rho}\right) \quad (33)$$

which is equivalent to condition (29), when ρ is constant everywhere (see Remark 14 for nonconstant ρ). After substitution of (32) in (33), the linear system of equations that determines $\alpha(x)$ is obtained as follows:

$$M(x) \alpha(x) = P(0) \quad (34)$$

where

$$M(x) = \sum_{j \in S_x^t} \omega(x_j, x) P\left(\frac{x_j - x}{\rho}\right) P^T\left(\frac{x_j - x}{\rho}\right) \quad (35)$$

Remark 14. The consistency conditions (29) and (33) are equivalent if the dilation parameter ρ is constant. When the dilation parameter varies at each particle, the same expression for $N_j(x)$ is used, namely equation (28), but the varying dilation parameter, ρ_j associated to particle x_j , is embedded in the definition of the weighting function; that is, equation (22) is modified as follows:

$$\omega(x_j, x) = \phi\left(\frac{x_j - x}{\rho_j}\right)$$

Note that a constant ρ is employed in the scaling of the polynomials P . The constant value ρ is typically chosen as the mean value of all the ρ_j . The consistency condition in this case is also (33). It also imposes the reproducibility of the polynomials in P .

This centered expression for the EFG shape functions can also be obtained with a discrete MLS development with the discrete centered scalar product

$$(f, g)_x = \sum_{j \in S_x^t} \omega(x_j, x) f\left(\frac{x_j - x}{\rho}\right) g\left(\frac{x_j - x}{\rho}\right) \quad (36)$$

The MLS development in this case is as follows: for fixed x , and for z near x , u is approximated as

$$u(z) \approx u^p(z, x) = P\left(\frac{z - x}{\rho}\right) c(x) \quad (37)$$

where c is obtained, as usual, through a least-squares fitting with the discrete centered scalar product (36).

2.2.6 The diffuse derivative

The centered MLS allows, with a proper definition of the polynomial basis P , the reinterpretation of the coefficients in $c(x)$ as approximations of u and its derivatives at the fixed point x .

The approximation of the derivative of u in each spatial direction is the corresponding derivative of u^p . This requires taking the derivative of (21), that is

$$\frac{\partial u}{\partial x} \simeq \frac{\partial u^p}{\partial x} = \frac{\partial \mathbf{P}^T}{\partial x} \mathbf{c} + \mathbf{P}^T \frac{\partial \mathbf{c}}{\partial x} \quad (38)$$

On one hand, the second term on the r.h.s. is not trivial. Derivatives of the coefficients \mathbf{c} require the resolution of a linear system of equations with the same matrix \mathbf{M} . As noted by Belytschko *et al.* (1996a), this is not an expensive task (see also Section 2.2.7). However, it requires the knowledge of the cloud of particles surrounding each point \mathbf{x} and, thus, it depends on the point where derivatives are evaluated.

On the other hand, the first term is easily evaluated. The derivative of the polynomials in \mathbf{P} is trivial and can be evaluated a priori, without knowledge of the cloud of particles surrounding each point \mathbf{x} .

Villon (1991) and Nayroles, Touzot and Villon (1992) propose the concept of diffuse derivative, which consists in approximating the derivative only with the first term on the r.h.s. of (38), namely,

$$\frac{\partial u^p}{\partial x} = \frac{\partial u^p}{\partial x} \bigg|_{z=x} = \frac{\partial \mathbf{P}^T}{\partial x} \bigg|_{z=x} \mathbf{c}(x) = \frac{\partial \mathbf{P}^T}{\partial x} \mathbf{c}(x)$$

From a computational cost point of view, this is an interesting alternative to (38). Moreover, the following proposition ensures convergence at the optimal rate of the diffuse derivative.

Proposition 2. *If u^p is an approximation to u with an order of consistency m (i.e. \mathbf{P} includes a complete basis of the subspace of polynomials of degree m) and ρ/h is constant, then*

$$\left\| \frac{\partial^k u^p}{\partial \mathbf{x}^k} - \frac{\partial^k u}{\partial \mathbf{x}^k} \right\|_{\infty} \leq C \frac{\rho^{m+1-|k|}}{(m+1)!} \quad \forall |k| = 0, \dots, m \quad (39)$$

where \mathbf{k} is a multiindex, $\mathbf{k} = (k_1, k_2, \dots, k_{nd})$ and $|\mathbf{k}| = k_1 + k_2 + \dots + k_{nd}$.

The proof can be found in Villon (1991) for 1D or Huerta, Vidal and Villon (2004b) for higher spatial dimensions. To clearly identify the coefficients of \mathbf{c} with the approximations of u and its derivatives, the basis in \mathbf{P} should be written appropriately. That is, each component of \mathbf{P} is $\mathbf{P}_\alpha(\mathbf{x}) = \xi^\alpha / \alpha!$ for $|\alpha| = 0, \dots, m$, where a standard multiindex notation is employed,

$$\mathbf{h}^\alpha := h_1^{\alpha_1} h_2^{\alpha_2} \dots h_{nd}^{\alpha_{nd}}; \quad \alpha! := \alpha_1! \alpha_2! \dots \alpha_{nd}!; \quad |\alpha| = \alpha_1 + \alpha_2 + \dots + \alpha_{nd}$$

Finally, it is important to emphasize the requirements that \mathbf{M} is regular and bounded (see Huerta, Fernández-Méndez and Díez, 2002; Fernández-Méndez and Huerta, 2002; Fernández-Méndez, Díez and Huerta, 2003).

Remark 15. For example, in 1D with consistency of order two, that is, $\mathbf{P}(\mathbf{x}) = [1, \rho x, (\rho x)^2/2]$, the expression (37) can be written as

$$u(z) \simeq u^p(z, x) = c_0(x) + c_1(x)(z - x) + c_2(x) \frac{(z - x)^2}{2} \quad (40)$$

Thus, taking the derivative with respect to z and imposing $z = x$,

$$u(x) \simeq c_0(x), \quad u'(x) \simeq c_1(x) \quad \text{and} \quad u''(x) \simeq c_2(x)$$

In fact, this strategy is the basis of the *diffuse element method*: the diffuse derivative is used in the weak form of the problem (see Nayroles, Touzot and Villon, 1992; Breikopf, Rassinoux and Villon, 2001). Moreover, the *generalized finite difference interpolation* or *meshless finite difference method* (see Orkisz, 1998) coincides also with this MLS development. The only difference between the generalized finite difference approximants and the EFG centered approximants is the definition of the set of neighboring particles S_x^c .

2.2.7 Direct evaluation of the derivatives

Another alternative to computing the derivatives is to fully derive the expression for the shape function (28) (see (38)), taking into account the dependencies given by equations (30) and (19). The details can be found in the references by Belytschko *et al.* (1996a,b).

For clarity, the 1D case is developed, $x \in \mathbb{R}$. For higher dimensions, the process is repeated for each component of \mathbf{x} . The derivative of the shape function (28) can be written as

$$\frac{dN_i}{dx}(x) = \omega(x_j, x) \mathbf{P}'(x_j) \frac{d\alpha}{dx}(x) + \frac{d\omega(x_j, x)}{dx} \mathbf{P}'(x_j) \alpha(x) \quad (41)$$

The derivative of the weighting function is easily obtained because $\omega(x_j, x)$ has usually known analytical expressions (see (22) and Section 2.1.2). The a priori nontrivial part is the evaluation of $d\alpha/dx$, but an expression for this vector can be obtained by implicit derivation of (30),

$$\mathbf{M}_x(x) \alpha(x) + \mathbf{M}(x) \frac{d\alpha}{dx}(x) = \frac{d\mathbf{P}}{dx}(x)$$

where matrix \mathbf{M}_x is the derivative of matrix \mathbf{M} ,

$$\mathbf{M}_x(x) = \sum_j \frac{d\omega(x_j, x)}{dx} \mathbf{P}(x_j) \mathbf{P}'(x_j)$$

which is trivial to compute. Thus, $d\alpha/dx$ is the solution of the linear system of equations

$$\mathbf{M}(x) \frac{d\alpha}{dx}(x) = \frac{d\mathbf{P}}{dx}(x) - \mathbf{M}_x(x) \alpha(x)$$

which represents another linear system of equations with the same matrix as in (30) (the factorization of \mathbf{M} can be reused) and a new independent term. Moreover, the product $\mathbf{M}_x(x) \alpha(x)$ can be computed in an efficient way as

$$\mathbf{M}_x(x) \alpha(x) = \sum_j \frac{d\omega(x_j, x)}{dx} \mathbf{P}(x_j) [\mathbf{P}'(x_j) \alpha(x)]$$

involving only vector operations.

In summary, the evaluation of the derivatives of the shape functions (see (41)) requires little extra computer cost and, moreover, higher-order derivatives can also be computed repeating the same process: it only depends on the regularity of $\omega(x_j, x)$. Obviously, the same development can be done for the centered and scaled approach defined in Section 2.2.5.

2.2.8 Partition of the unity methods

The set of MLS approximation functions can be viewed as a partition of unity: the approximation verifies, at least, the 0-order consistency condition (reproducibility of the constant polynomial $p(x) = 1$)

$$\sum_i N_i \cdot 1 = 1$$

This viewpoint leads to new approximations for meshfree methods. Based on the *partition of the unity finite element method* (PUFEM) proposed by Babuska, Banerjee and Osborn (2003), Babuska and Melenk (1995), and Duarte and Oden (1996b) use the concept of partition of unity to construct approximations with consistency of order $k \geq 1$. They call their method *h-p clouds*. The approximation is

$$u(\mathbf{x}) \simeq u^p(\mathbf{x}) = \sum_i N_i(\mathbf{x}) u_i + \sum_{i=1}^{n_l} b_{i1} [N_i(\mathbf{x}) q_{i1}(\mathbf{x})]$$

where $N_i(\mathbf{x})$ are the MLS approximation functions, q_{i1} are n_l polynomials of degree greater than k associated to

each particle x_j , and u_j, b_{i1} are coefficients to determine. Note that the polynomials $q_{i1}(\mathbf{x})$ increase the order of the approximation space. These polynomials can be different from particle to particle, thus facilitating the *hp*-adaptivity.

Remark 16. As commented in Belytschko *et al.* (1996b), the concept of an extrinsic basis, $q_{i1}(\mathbf{x})$, is essential for obtaining *p*-adaptivity. In MLS approximations, the intrinsic basis \mathbf{P} cannot vary from particle to particle without introducing a discontinuity.

3 DISCRETIZATION OF PARTIAL DIFFERENTIAL EQUATIONS

All of the approximation functions described in the previous section can be used in the solution of a partial differential equation (PDE) boundary value problem. Usually SPH methods are combined with a collocation or point integration technique, while the approximations based on MLS are usually combined with a Galerkin discretization.

A large number of published methods with minor differences exist. Probably, the best known, among those using MLS approximation, are the *meshless (generalized) finite difference method* (MFDM) developed by Liszka and Orkisz (1980), the DEM by Nayroles, Touzot and Villon (1992), the *element-free Galerkin* (EFG) method by Belytschko, Lu and Gu (1994), the RKPM by Liu, Jun and Zhang (1995b), the *meshless local Petrov–Galerkin* (MLPG) by Zhu and Atluri (1998), the *corrected smooth particle hydrodynamics* (CSPH) by Bonet and Lok (1999), and the *finite point method* (FPM) by Oñate and Idelsohn (1998). Table 1 classifies these methods depending on the evaluation of the derivatives (see Sections 2.2.6 and 2.2.7) and how the PDE is solved (Galerkin, Petrov–Galerkin, point collocation).

Partition of unity methods can be implemented with Galerkin and Petrov–Galerkin methods. For instance, the *h-p clouds* by Duarte and Oden (1996b) uses a Galerkin

Table 1. Classification of MLS based meshfree methods for PDE's.

| | | Evaluation of derivatives | |
|----------|---------------------|---------------------------|-------------|
| | | Direct | Diffuse |
| Galerkin | Gauss quadrature | EFG RKPM | MFDM DEM |
| | Particle quadrature | CSPH* | |
| | Petrov–Galerkin | MLPG | |
| | Point collocation | | FPM |

*Direct and global evaluation of derivatives.

weak form with accurate integration, while the *finite spheres* by De and Bathe (2000) uses Shepard functions (with circular/spherical support) enriched with polynomials and a Galerkin weak form with specific quadratures for spherical domains (and particular/adapted quadratures near the boundary); almost identical to *h-p clouds* (see Duarte and Oden, 1996b).

In order to discuss some of these methods in more detail, the model boundary value problem

$$\Delta u - u = -f \quad \text{in } \Omega \quad (42a)$$

$$u = u_D \quad \text{on } \Gamma_D \quad (42b)$$

$$\frac{\partial u}{\partial n} = g_N \quad \text{on } \Gamma_N \quad (42c)$$

is considered, where Δ is the Laplace operator in 2D, $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$, \mathbf{n} is the unitary outward normal vector on $\partial\Omega$, $\partial/\partial n$ is the normal derivative, $\partial/\partial n = n_1 \partial/\partial x + n_2 \partial/\partial y$, $\Gamma_D \cup \Gamma_N = \partial\Omega$, and f , u_D , and g_N are known.

Remark 17. A major issue in the implementation of meshfree methods is the identification (localization) of the neighboring particles. That is, given a point \mathbf{x} in the domain, identify which particles have a nonzero shape function at this point, that is, determine the x_i such that $\phi((x_j - x)/\rho) \neq 0$ to define S_x^2 .

There are several options. The usual ones consist in determining the closest particles to the point of interest (see Randles and Libersky, 2000) or to use a nonconforming regular mesh of squares or cubes (cells) parallel to the Cartesian axes. For every cell, the indices of the particles inside the cell are stored. The regular structure of the cell mesh allows one to find, given a point \mathbf{x} , the cell where \mathbf{x} is located and, thus, determine the neighboring particles just looking in the neighboring cells. Fast algorithms for finding neighboring particles can be found in the book by Schweitzer (2003).

3.1 Collocation methods

Consider an approximation, based on a set of particles $\{x_i\}$, of the form

$$u(\mathbf{x}) \simeq u^p(\mathbf{x}) = \sum_i u_i N_i(\mathbf{x})$$

The shape functions $N_i(\mathbf{x})$ can be SPH shape functions (Section 2.1.1) or MLS shape functions (Section 2.2), and u_i are coefficients to be determined.

In collocation methods (see Oñate and Idelsohn, 1998 or Aluru, 2000), the PDE (42a) is imposed at each particle

in the interior of the domain Ω , the boundary conditions (42b) and (42c) are imposed at each particle of the corresponding boundary. In the case of the model problem, this leads to the linear system of equations for the coefficients u_i :

$$\begin{cases} \sum_i u_i [\Delta N_i(x_j) - N_i(x_j)] = -f(x_j) & \forall x_j \in \Omega \\ \sum_i u_i N_i(x_j) = u_D(x_j) & \forall x_j \in \Gamma_D \\ \sum_i u_i \frac{\partial N_i}{\partial n}(x_j) = g_N(x_j) & \forall x_j \in \Gamma_N \end{cases}$$

Note that the shape functions must be C^2 , and thus, a C^2 window function must be used. In this case, the solution at particle x_j is approximated by

$$u(x_j) \simeq u^p(x_j) = \sum_i u_i N_i(x_j)$$

which in general differs from the coefficient u_j (see Remark 1).

In the context of the renormalized meshless derivative (see Vila, 1999), the coefficient u_j is considered as the approximation at the particle x_j and only the derivative of the solution is approximated through the RMD (see (9)). Thus, the linear system to be solved becomes

$$\begin{cases} \sum_i u_i \Delta N_i(x_j) - u_j = -f(x_j) & \forall x_j \in \Omega \\ u_j = u_D(x_j) & \forall x_j \in \Gamma_D \\ \sum_i u_i \frac{\partial N_i}{\partial n}(x_j) = g_N(x_j) & \forall x_j \in \Gamma_N \end{cases}$$

Both possibilities are slightly different from the SPH method by Monaghan (1988) or from SPH methods based on particle integration techniques (see Bonet and Kulasegaram, 2000).

3.2 Methods based on a Galerkin weak form

The meshfree shape functions can also be used in the discretization of the weak integral form of the boundary value problem. For the model problem (42), the Bubnov-Galerkin weak form (also used in the finite element method (FEM)) is

$$\begin{aligned} \int_{\Omega} \nabla v \nabla u \, d\Omega + \int_{\Omega} v u \, d\Omega \\ = \int_{\Omega} v f \, d\Omega + \int_{\Gamma_N} v g_N \, d\Gamma \quad \forall v \end{aligned}$$

where v vanishes at Γ_D and $u = u_D$ at Γ_D . However, this weak form can not be directly discretized with a standard meshfree approximation. The shape functions do not verify the Kronecker delta property (see Remark 1) and, thus, it is difficult to select v such that, $v = 0$ at Γ_D and to impose that $u^p = u_D$ at Γ_D . Specific techniques are needed in order to impose Dirichlet boundary conditions. Section 3.3 describes the treatment of essential boundary conditions in meshfree methods.

An important issue in the implementation of a meshfree method with a weak form is the evaluation of integrals.

Several possibilities can be considered to evaluate integrals in the weak form: (1) the particles can be the quadrature points (*particle integration*), (2) a regular cell structure (for instance, the same used for the localization of particles) can be used with numerical quadrature in each cell (*cell integration*), or (3) a, not necessary regular, *background mesh* can be used to compute integrals. The first possibility (*particle integration*) is the fastest, but as in collocation methods, it can result in rank deficiency. Bonet and Kulasegaram (2000) propose a global correction to obtain accurate and stable results with particle integration. The other two possibilities present the disadvantage that, since shape functions and their derivatives are not polynomials, the number of integration points leads to high computational costs. Nevertheless, the cell structure or the coarse background mesh, which do not need to be conforming, are easily generated and these techniques ensure an accurate approximation of the integrals (see Chapter 4, this Volume).

3.3 Essential boundary conditions

Many specific techniques have been developed in the recent years in order to impose essential boundary conditions in meshfree methods. Some possibilities are (1) Lagrange multipliers (Belytschko, Lu and Gu, 1994), (2) modified variational principles (Belytschko, Lu and Gu, 1994), (3) penalty methods (Zhu and Atluri, 1998; Bonet and Kulasegaram, 2000), (4) perturbed Lagrangian (Chiu and Moran, 1995), (5) coupling to finite elements (Belytschko, Organ and Krongauz, 1995; Huerta and Fernández-Méndez, 2000a; Wagner and Liu, 2001), or (6) modified shape functions (Gosz and Liu, 1996; Günter and Liu, 1998; Wagner and Liu, 2000) among others.

The first attempts to define shape functions with the 'delta property' along the boundary (see Gosz and Liu, 1996), namely, $N_i(x_j) = \delta_{ij}$ for all x_j in Γ_D , have serious difficulties for complex domains and for the integration of the weak forms.

In the recent years, mixed interpolations that combine finite elements with meshfree methods have been

developed. Mixed interpolations can be quite effective for imposing essential boundary conditions. The idea is to use one or two layers of finite elements next to the Dirichlet boundary and use a meshfree approximation in the rest of the domain. Thus, the essential boundary conditions can be imposed as in standard finite elements. In Belytschko, Organ and Krongauz (1995), a mixed interpolation is defined in the transition area (from the finite elements region to the particles region). This mixed interpolation requires the substitution of finite element nodes by particles and the definition of ramp functions. Thus, the transition is of the size of one element and the interpolation is linear. Following this idea, Huerta and Fernández-Méndez (2000a) propose a more general mixed interpolation, for any order of interpolation with no need for ramp functions and no substitution of nodes by particles. This is done preserving consistency and continuity of the solution. Figure 8 shows an example of this mixed interpolation in 1D: two finite element nodes are considered at the boundary of the domain, with their corresponding shape functions with a dashed line, and the meshfree shape functions are modified in order to preserve consistency, with a solid line. The details of this method are presented in Section 6.

3.3.1 Methods based on a modification of the weak form

For the sake of clarity, the following model problem is considered

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = u_D & \text{on } \Gamma_D \\ \nabla u \cdot \mathbf{n} = g_N & \text{on } \Gamma_N \end{cases} \quad (43)$$

where $\Gamma_D \cup \Gamma_N = \partial\Omega$, $\Gamma_D \cap \Gamma_N = \emptyset$ and \mathbf{n} is the outward unit normal on $\partial\Omega$. The generalization of the following developments to other PDEs is straightforward.

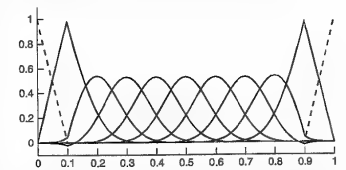


Figure 8. Mixed interpolation with linear finite element nodes near the boundary and particles in the interior of the domain, with $\rho/h = 3.2$, cubic spline and linear consistency in all the domain.

The weak problem form of (43) is 'find $u \in \mathcal{H}^1(\Omega)$ such that $u = u_D$ on Γ_D and

$$\int_{\Omega} \nabla v \cdot \nabla u \, d\Omega - \int_{\Gamma_D} v \nabla u \cdot \mathbf{n} \, d\Gamma = \int_{\Omega} v f \, d\Omega + \int_{\Gamma_N} v g_N \, d\Gamma \quad (44)$$

for all $v \in \mathcal{H}_0^1(\Omega)'$. In the finite element method, the interpolation of u can easily be forced to verify the essential boundary condition and the test functions v can be chosen such that $v = 0$ on Γ_D (see Remark 18), leading to the following weak form: 'find $u \in \mathcal{H}^1(\Omega)$ such that $u = u_D$ on Γ_D and

$$\int_{\Omega} \nabla v \cdot \nabla u \, d\Omega = \int_{\Omega} v f \, d\Omega + \int_{\Gamma_N} v g_N \, d\Gamma \quad (45)$$

for all $v \in \mathcal{H}_0^1(\Omega)'$, where $\mathcal{H}_0^1(\Omega) = \{v \in \mathcal{H}^1(\Omega) \mid v = 0 \text{ on } \Gamma_D\}$.

Remark 18. In the FEM, or in the context of the continuous blending method discussed in Section 6, the approximation can be written as

$$u(x) \simeq \sum_{j \in \mathcal{B}} u_j N_j(x) + \psi(x) \quad (46)$$

where $N_j(x)$ denotes the shape functions, $\psi(x) = \sum_{i \in \mathcal{B}} u_D(x_i) N_i(x)$, and \mathcal{B} is the set of indexes of all nodes on the essential boundary. Note that, because of the Kronecker delta property of the shape functions for $i \in \mathcal{B}$ and the fact that $N_i \in \mathcal{H}_0^1(\Omega)$ for $i \notin \mathcal{B}$, the approximation defined by (46) verifies $u = u_D$ at the nodes on the essential boundary. Therefore, approximation (46) and $v = N_i$, for $i \notin \mathcal{B}$, can be considered for the discretization of the weak form (45). Under these circumstances, the system of equations becomes

$$\mathbf{K} \mathbf{u} = \mathbf{f} \quad (47)$$

where

$$K_{ij} = \int_{\Omega} \nabla N_i \cdot \nabla N_j \, d\Omega$$

$$f_i = \int_{\Omega} N_i f \, d\Omega + \int_{\Gamma_D} N_i \psi \, d\Omega + \int_{\Gamma_N} N_i g_N \, d\Gamma \quad (48)$$

and \mathbf{u} is the vector of coefficients u_i .

However, for standard meshfree approximation, the shape functions do not verify the Kronecker delta property and $N_i \notin \mathcal{H}_0^1(\Omega)$ for $i \notin \mathcal{B}$. Therefore, imposing $u = u_D$ and $v = 0$ on Γ_D is not as straightforward as in finite elements

or as in the blending method (Belytschko, Organ and Krongauz, 1995), and the weak form defined by (45) cannot be used. The most popular methods that modify the weak form to overcome this problem are the Lagrange multiplier method, the penalty method, and Nitsche's method.

3.3.2 Lagrange multiplier method

The solution of problem (43) can also be obtained as the solution of a minimization problem with constraints: ' u minimizes the energy functional

$$\Pi(v) = \frac{1}{2} \int_{\Omega} \nabla v \cdot \nabla v \, d\Omega - \int_{\Omega} v f \, d\Omega - \int_{\Gamma_N} v g_N \, d\Gamma \quad (49)$$

and verifies the essential boundary conditions.' That is,

$$u = \arg \min_{\substack{v \in \mathcal{H}^1(\Omega) \\ v = u_D \text{ on } \Gamma_D}} \Pi(v) \quad (50)$$

With the use of a Lagrange multiplier, $\lambda(x)$, this minimization problem can also be written as

$$(u, \lambda) = \arg \min_{u \in \mathcal{H}^1(\Omega)} \max_{\gamma \in \mathcal{H}^{-1/2}(\Gamma_D)} \Pi(v) + \int_{\Gamma_D} \gamma (v - u_D) \, d\Gamma$$

This min-max problem leads to the following weak form with Lagrange multiplier, 'find $u \in \mathcal{H}^1(\Omega)$ and $\lambda \in \mathcal{H}^{-1/2}(\Gamma_D)$ such that

$$\int_{\Omega} \nabla v \cdot \nabla u \, d\Omega + \int_{\Gamma_D} v \lambda \, d\Gamma = \int_{\Omega} v f \, d\Omega + \int_{\Gamma_N} v g_N \, d\Gamma, \quad \forall v \in \mathcal{H}^1(\Omega) \quad (51a)$$

$$\int_{\Gamma_D} \gamma (u - u_D) \, d\Gamma = 0, \quad \forall \gamma \in \mathcal{H}^{-1/2}(\Gamma_D)' \quad (51b)$$

Remark 19. Equation (51b) imposes the essential boundary condition, $u = u_D$ on Γ_D , in weak form.

Remark 20. The physical interpretation of the Lagrange multiplier can be seen by simple comparison of equations (51a) and (44): the Lagrange multiplier corresponds to the flux (traction in a mechanical problem) along the essential boundary, $\lambda = -\nabla u \cdot \mathbf{n}$.

Considering now the approximation $u(x) \simeq \sum N_i(x) u_i$ with meshfree shape functions N_i and an interpolation for λ with a set of boundary functions $\{N_i^L(x)\}_{i=1}^{\ell}$,

$$\lambda(x) \simeq \sum_{i=1}^{\ell} \lambda_i N_i^L(x) \quad \text{for } x \in \Gamma_D \quad (52)$$

the discretization of (51) leads to the system of equations

$$\begin{pmatrix} \mathbf{K} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{b} \end{pmatrix} \quad (53)$$

where \mathbf{K} and \mathbf{f} are already defined in (48) (use $\psi = 0$), $\boldsymbol{\lambda}$ is the vector of coefficients λ_i , and

$$A_{ij} = \int_{\Gamma_D} N_i^L N_j \, d\Gamma, \quad b_i = \int_{\Gamma_D} N_i^L u_D \, d\Gamma$$

There are several possibilities for the choice of the interpolation space for the Lagrange multiplier λ . Some of them are (1) a finite element interpolation on the essential boundary, (2) a meshfree approximation on the essential boundary, or (3) the same shape functions used in the interpolation of u restricted along Γ_D , that is, $N_i^L = N_i$ for i such that $N_i|_{\Gamma_D} \neq 0$. However, the most popular choice is the point collocation method. This method corresponds to $N_i^L(x) = \delta(x - x_i^L)$, where $\{x_i^L\}_{i=1}^{\ell}$ is a set of points along Γ_D and δ is the Dirac delta function. In that case, by substitution of $\gamma(x) = \delta(x - x_i^L)$, equation (51b) corresponds to

$$u(x_i^L) = u_D(x_i^L), \quad \text{for } i = 1, \dots, \ell$$

That is, $A_{ij} = N_j(x_i^L)$, $b_i = u_D(x_i^L)$, and each equation of $\mathbf{A} \mathbf{u} = \mathbf{b}$ in (53) corresponds to the enforcement of the prescribed value at one collocation point, namely, x_i^L .

Remark 21. The system of equations (53) can also be derived from the minimization in \mathbb{R}^{n+m} of the discrete version of the energy functional (49) subject to the constraints corresponding to the essential boundary conditions, $\mathbf{A} \mathbf{u} = \mathbf{b}$. In fact, there is no need to know the weak form with Lagrange multiplier, it is sufficient to define the discrete energy functional and the restrictions due to the boundary conditions in order to determine the system of equations.

Therefore, the Lagrange multiplier method is, in principle, general and easily applicable to all kind of problems. However, the main disadvantages of the Lagrange multiplier method are

1. The dimension of the resulting system of equations is increased.
2. Even for \mathbf{K} symmetric and semi-positive definite, the global matrix in (53) is symmetric but it is no longer positive definite. Therefore, standard linear solvers for symmetric and positive definite matrices cannot be used.
3. More crucial is the fact that the system (53) and the weak problem (51) induce a saddle-point problem, which precludes an arbitrary choice of the interpolation space for u and λ . The resolution of the multiplier

λ field must be fine enough in order to obtain an acceptable solution, but the system of equations will be singular if the resolution of the Lagrange multiplier λ field is too fine. In fact, the interpolation spaces for the Lagrange multiplier λ and for the principal unknown u must verify an inf-sup condition, known as the Babuska-Brezzi stability condition, in order to ensure the convergence of the approximation (see Babuska, 1973a or Brezzi, 1974 for details).

The first two disadvantages can be neglected in view of the versatility and simplicity of the method. However, while in the FEM, it is trivial to choose the approximation for the Lagrange multiplier in order to verify the Babuska-Brezzi stability condition and to impose accurate essential boundary conditions, this choice is not trivial for meshfree methods. In fact, in meshfree methods, the choice of an appropriate interpolation for the Lagrange multiplier can be a serious problem in particular situations.

These properties are observed in the resolution of the 2D linear elasticity problem represented in Figure 9 where the solution obtained with a regular mesh of 30×30 biquadratic finite elements is also shown. The distance between particles is $h = 1/6$ and a finer mesh is used for the representation of the solution.

Figure 10 shows the solution obtained for the Lagrange multiplier method. The prescribed displacement is imposed at some collocation points at the essential boundary (marked with black squares). Three possible distributions for the collocation points are considered. In the first one, the collocation points correspond to the particles located at the essential boundary. The prescribed displacement is exactly imposed at the collocation points, but not along the rest of the essential boundary. Note that the displacement field is not accurate because of the smoothness of the meshfree approximation. But if the number of collocation points is too large, the inf-sup condition is no longer verified and the system stiffness matrix is singular. This is the case of discretization (c), which corresponds to double the

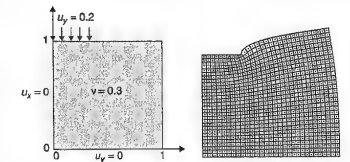


Figure 9. Problem statement and solution with 30×30 biquadratic finite elements (61 \times 61 nodes).

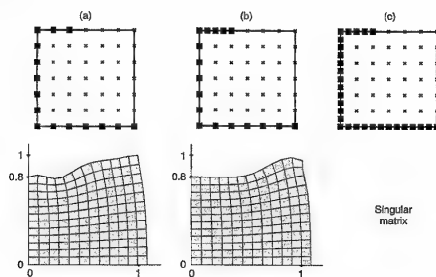


Figure 10. Solution with Lagrange multipliers for three possible distributions of collocation points (■) and 7×7 particles.

density of collocation points along the essential boundary. In this example, the choice of a proper interpolation for the Lagrange multiplier is not trivial. Option (b) represents a distribution of collocation points that imposes the prescribed displacements in a correct manner and, at the same time, leads to a regular matrix. Similar results are obtained if the Lagrange multiplier is interpolated with boundary linear finite elements (see Fernández-Méndez and Huerta, 2004).

Therefore, although imposing boundary constraints is straightforward with the Lagrange multiplier method, the applicability of this method in particular cases is impaired due to the difficulty in the selection of a proper interpolation space for the Lagrange multiplier. It is important to note that the choice of the interpolation space can be even more complicated for an irregular distribution of particles (see Chapter 9, this Volume).

3.3.3 Penalty method

The minimization problem with constraints defined by (50) can also be solved with the use of a penalty parameter. That is,

$$u = \arg \min_{v \in \mathcal{H}^1(\Omega)} \Pi(v) + \frac{1}{2} \beta \int_{\Gamma_D} (v - u_D)^2 d\Gamma \quad (54)$$

The penalty parameter β is a positive scalar constant that must be large enough to accurately impose the essential boundary condition. The minimization problem (54) leads to the following weak form: 'find $u \in \mathcal{H}^1(\Omega)$ such that

$$\int_{\Omega} \nabla v \cdot \nabla u \, d\Omega + \beta \int_{\Gamma_D} v u \, d\Gamma = \int_{\Omega} v f \, d\Omega$$

$$+ \int_{\Gamma_N} v g_N \, d\Gamma + \beta \int_{\Gamma_D} v u_D \, d\Gamma \quad (55)$$

for all $v \in \mathcal{H}^1(\Omega)$. The discretization of this weak form leads to the system of equations

$$(\mathbf{K} + \beta \mathbf{M}^p) \mathbf{u} = \mathbf{f} + \beta \mathbf{f}^p \quad (56)$$

where \mathbf{K} and \mathbf{f} are defined in (48) (use $\psi = 0$) and

$$\mathbf{M}_{ij}^p = \int_{\Gamma_D} N_i N_j \, d\Gamma, \quad \mathbf{f}_i^p = \int_{\Gamma_D} N_i u_D \, d\Gamma$$

Remark 22. The penalty method can also be obtained from the minimization of the discrete version of the energy functional in $\mathbb{R}^{n_{tot}}$, subjected to the constraints corresponding to the essential boundary condition $\mathbf{A}\mathbf{u} = \mathbf{b}$.

Like the Lagrange multiplier method, the penalty method is easily applicable to a wide range of problems. The penalty method presents two clear advantages: (1) the dimension of the system is not increased and (2) the matrix in the resulting system (see equation (56)) is symmetric and positive definite, provided that \mathbf{K} is symmetric and β is large enough.

However, the penalty method also has two important drawbacks: the Dirichlet boundary condition is weakly imposed (the parameter β controls how well the essential boundary condition is met) and the matrix in (56) is often poorly conditioned (the condition number increases with β).

A general theorem on the convergence of the penalty method and the choice of the penalty parameter β can be found in Babuska (1973b) and Babuska, Banerjee and

Osborn (2002b). For an interpolation with consistency of order p and discretization measure h (i.e. the characteristic element size in finite elements or the characteristic distance between particles in a meshfree method), the best error estimate obtained by Babuska (1973b) gives a rate of convergence of order $h^{(2p+1)/3}$ in the energy norm, provided that the penalty β is taken to be of order $h^{-(2p+1)/3}$. In the linear case, it corresponds to the optimal rate of convergence in the energy norm. For order $p \geq 2$, the lack of optimality in the rate of convergence is a direct consequence of the lack of consistency of the weak formulation (see Arnold *et al.*, 2001/02 and Remark 23).

These properties can be observed in the following 2D Laplace problem:

$$\begin{cases} \Delta u = 0 & (x, y) \in]0, 1[\times]0, 1[\\ u(x, 0) = \sin(\pi x) \\ u(x, 1) = u(0, y) = u(1, y) = 0 \end{cases}$$

with analytical solution (see Wagner and Liu, 2000),

$$u(x, y) = [\cosh(\pi y) - \coth(\pi y) \sinh(\pi y)] \sin(\pi x)$$

A distribution of 7×7 particles is considered, that is, the distance between particles is $h = 1/6$.

Figure 11 shows the solution for increasing values of the penalty parameter β . The penalty parameter must be large enough, $\beta \geq 10^3$, in order to impose the boundary condition in an accurate manner. Figure 12 shows convergence curves for different choices of the penalty parameter. The penalty method converges with a rate close to 2 in the L^2 norm if the penalty parameter β is proportional to h^{-2} . If the penalty parameter is constant, or proportional to h^{-1} , the

boundary error dominates and the optimal convergence rate is lost as h goes to zero.

Figure 12 also shows the matrix condition number for increasing values of the penalty parameter, for a distribution of 11×11 and 21×21 particles. The condition number grows linearly with the penalty parameter. Note that, for instance, for a discretization of 21×21 particles, a reasonable value for the penalty parameter is $\beta = 10^6$, which corresponds to a condition number near 10^{12} . Obviously, the situation gets worse for denser discretizations, which need larger penalty parameters. The ill-conditioning of the matrix reduces the applicability of the penalty method.

3.3.4 Nitsche's method

Nitsche's weak form for problem (43) is

$$\begin{aligned} \int_{\Omega} \nabla v \cdot \nabla u \, d\Omega - \int_{\Gamma_D} v \nabla u \cdot \mathbf{n} \, d\Gamma - \int_{\Gamma_D} \nabla v \cdot \mathbf{n} u \, d\Gamma \\ + \beta \int_{\Gamma_D} v u \, d\Gamma = \int_{\Omega} v f \, d\Omega + \int_{\Gamma_N} v g_N \, d\Gamma \\ - \int_{\Gamma_D} \nabla v \cdot \mathbf{n} u_D \, d\Gamma + \beta \int_{\Gamma_D} v u_D \, d\Gamma \end{aligned} \quad (57)$$

where β is a positive constant scalar parameter (see Arnold *et al.*, 2001/02; Nitsche, 1970).

Comparing with the weak form defined by (44), the new terms in the l.h.s. of (57) are $\int_{\Gamma_D} u \nabla v \cdot \mathbf{n} \, d\Gamma$, which recovers the symmetry of the bilinear form, and $\beta \int_{\Gamma_D} v u \, d\Gamma$, which ensures the coercivity of the bilinear form (i.e. the matrix corresponding to its discretization is positive definite), provided that β is large enough. The new terms

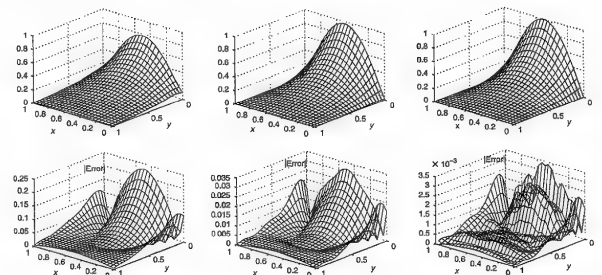


Figure 11. Penalty method solution (top) and error (bottom) for $\beta = 10$ (left), $\beta = 100$ (center) and $\beta = 10^3$ (right).

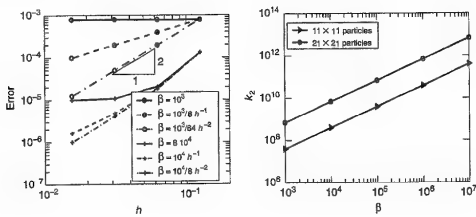


Figure 12. Evolution of the $L^2(\Omega)$ error norm for the penalty method and matrix condition number.

in the r.h.s. are added to ensure consistency of the weak form.

The discretization of the Nitsche's weak form leads to a system of equations with the same size as \mathbf{K} and whose system matrix is symmetric and positive definite, provided that \mathbf{K} is symmetric and β is large enough. Although, as in the penalty method, the condition number of this matrix increases with parameter β , in practice not very large values are needed in order to ensure convergence and a proper implementation of the boundary condition.

Remark 23. Nitsche's method can be interpreted as a consistent improvement of the penalty method. The penalty weak form (55) is not consistent, in the sense that the solution of (43) does not verify the penalty weak form for trial test functions that do not vanish at Γ_D (see Arnold *et al.*, 2001/02). Nitsche's weak form keeps the term $\int_{\Gamma_D} v \nabla u \cdot \mathbf{n} \, d\Gamma$ from the consistent weak form (44) and includes new terms maintaining the consistency.

The only problem of Nitsche's method is the deduction of the weak form. The generalization of the implementation for other problems is not as straightforward as for the method of Lagrange multipliers or for the penalty method. The weak form and the choice of the parameter β depends not only on the partial differential equation, but also on the essential boundary condition to be prescribed. Nitsche's method applied to other problems is discussed by Nitsche (1970), by Becker (2002) for the Navier-Stokes problem, by Freund and Stenberg (1995) for the Stokes problem, by Hansbo and Larson (2002) for elasticity problems.

Regarding the choice of the parameter, Nitsche proved that if β is taken as $\beta = \alpha/h$, where α is a large enough constant and h denotes the characteristic discretization measure, then the discrete solution converges to the exact solution with optimal order in H^1 and L^2 norms. Moreover,

for model problem (43) with Dirichlet boundary conditions, $\Gamma_D = \partial\Omega$, a value for constant α can be determined taking into account that convergence is ensured if $\beta > 2C^2$, where C is a positive constant such that $\|\nabla v \cdot \mathbf{n}\|_{L^2(\partial\Omega)} \leq C\|\nabla v\|_{L^2(\Omega)}$ for all v in the chosen interpolation space. This condition ensures the coercivity of the bilinear form in the interpolation space. Griebel and Schweitzer (2000) propose the estimation of the constant C as the maximum eigenvalue of the generalized eigenvalue problem,

$$A\mathbf{v} = \lambda B\mathbf{v} \quad (58)$$

where

$$A_{ij} = \int_{\Omega} (\nabla N_i \cdot \mathbf{n})(\nabla N_j \cdot \mathbf{n}) \, d\Gamma,$$

$$B_{ij} = \int_{\Omega} \nabla N_i \cdot \nabla N_j \, d\Omega$$

The problem described in Figure 9 can be solved by Nitsche's method for different values of β (see Figure 13). Note that the modification of the weak form is not trivial in this case (see Fernández-Méndez and Huerta, 2004). The major advantage of Nitsche's method is that scalar parameter β need not be as large as in the penalty method, and avoids the need to meet the Babuska-Brezzi condition for the interpolation space for the Lagrange multiplier.

3.4 Incompressibility and volumetric locking in meshfree methods

Locking in finite elements has been a major concern since its early developments; it is of particular concern for non-linear materials (Belytschko, Liu and Moran, 2000). It appears because poor numerical interpolation leads to an overconstrained system. Locking of standard finite elements

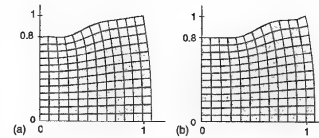


Figure 13. Nitsche's solution, 7×7 distribution of particles, for $\beta = 100$ (a) and $\beta = 10^4$ (b).

has been extensively studied. It is well known that bilinear finite elements lock in some problems and that biquadratic elements have a better behavior (Hughes, 2000). Moreover, locking has also been studied for increasing polynomial degrees in the context of an hp adaptive strategy (see Suri, 1996).

For instance, let's consider a linear elastic isotropic material under plane strain conditions and small deformations, namely, $\nabla^s \mathbf{u}$, where \mathbf{u} is the displacement and ∇^s the symmetric gradient, that is, $\nabla^s = (1/2)(\nabla^T + \nabla)$. Dirichlet boundary conditions are imposed on Γ_D , a traction \mathbf{h} is prescribed along the Neumann boundary Γ_N and there is a body force \mathbf{f} . Thus, the problem that needs to be solved may be stated as: solve for $\mathbf{u} \in [H_{0,\Gamma_D}^1]^2$ such that

$$\frac{E}{1+\nu} \int_{\Omega} \nabla^s \mathbf{v} : \nabla^s \mathbf{u} \, d\Omega$$

$$+ \frac{E\nu}{(1+\nu)(1-2\nu)} \int_{\Omega} (\nabla \cdot \mathbf{v})(\nabla \cdot \mathbf{u}) \, d\Omega$$

$$= \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\Omega + \int_{\Gamma_N} \mathbf{h} \cdot \mathbf{v} \, d\Gamma \quad \forall \mathbf{v} \in [H_{0,\Gamma_D}^1]^2 \quad (59)$$

In this equation, the standard vector subspaces of H^1 are employed for the solution \mathbf{u} , $[H_{0,\Gamma_D}^1]^2 := \{\mathbf{u} \in [H^1]^2 \mid \mathbf{u} = \mathbf{u}_D \text{ on } \Gamma_D\}$ (Dirichlet conditions, \mathbf{u}_D , are automatically satisfied) and for the test functions \mathbf{v} , $[H_{0,\Gamma_D}^1]^2 := \{\mathbf{v} \in [H^1]^2 \mid \mathbf{v} = \mathbf{0} \text{ on } \Gamma_D\}$, (zero values are imposed along Γ_D).

This equation, as discussed by Suri (1996), shows the inherent difficulties of the incompressible limit. The standard a priori error estimate emanating from (59) and based on the energy norm may become unbounded for values of ν close to 0.5. In fact, in order to have finite values of the energy norm, the divergence-free condition must be enforced in the continuum case, that is, $\nabla \cdot \mathbf{u} = 0$ for $\mathbf{u} \in [H_{0,\Gamma_D}^1]^2$, and also in the finite dimensional approximation space. In fact, locking will occur when the approximation space is not rich enough for the approximation to verify the divergence-free condition.

In fluids, incompressibility is directly the concern. Accurate and efficient modelling of incompressible flows is an

important issue in finite elements. The continuity equation for an incompressible fluid takes a peculiar form. It consists of a constraint on the velocity field that must be divergence-free. Then, the pressure has to be considered as a variable not related to any constitutive equation. Its presence in the momentum equation has the purpose of introducing an additional degree of freedom needed to satisfy the incompressibility constraint. The role of the pressure variable is thus to adjust itself instantaneously in order to satisfy the condition of divergence-free velocity. That is, the pressure is acting as a Lagrange multiplier of the incompressibility constraint and thus there is a coupling between the velocity and the pressure unknowns.

Various formulations have been proposed for incompressible flow (Girault and Raviart, 1986; Gresho and Sanli, 2000; Gunzburger, 1989; Pironneau, 1989; Quartapelle, 1993; Quarteroni and Valli, 1994; Temam, 2001; Donca and Huerta, 2003). Mixed finite elements present numerical difficulties caused by the saddle-point nature of the resulting variational problem. Solvability of the problem depends on a proper choice of finite element spaces for velocity and pressure. They must satisfy a compatibility condition, the so-called *Ladyzhenskaya-Babuska-Brezzi* (LBB or inf-sup) condition. If this is not the case, alternative formulations (usually depending on a numerical parameter) are devised to circumvent the LBB condition and enable the use of velocity-pressure pairs that are unstable in the standard Galerkin formulation.

Incompressibility in meshfree methods is still an open topic. Even recently, it was claimed (Belytschko, Lu and Gu, 1994; Zhu and Atluri, 1998) that meshless methods do not exhibit volumetric locking. Now it is clear that this is not true, as shown in a study of the element-free Galerkin (EFG) method by Dolbow and Belytschko (1999). Moreover, several authors claim that by increasing the dilation parameter, locking phenomena in meshfree methods can be suppressed or at least attenuated (Askes, Borst and Heeres, 1999; Dolbow and Belytschko, 1999; Chen *et al.*, 2000). Huerta and Fernández-Méndez (2001) clarify this issue and determine the influence of the dilation parameter on the locking behavior of EFG near the incompressible limit by a modal analysis. The major conclusions are

1. The number of nonphysical locking modes is independent of the ratio ρ/h .
2. An increase of the dilation parameter decreases the eigenvalue (amount of energy) in the locking mode and attenuates, but does not suppress volumetric locking (in the incompressible limit the energy will remain unbounded).
3. An increase in the order of consistency decreases the number of nonphysical locking modes.

4. The decrease in the number of nonphysical locking modes is slower than in finite elements. Thus EFG will not improve the properties of the FEM (from a volumetric locking viewpoint) for p or h - p refinement. However, for practical purposes and as in finite elements, in EFG an h - p strategy will also suppress locking.

The remedies proposed in the literature are, in general, extensions of the methods developed for finite elements. For instance, Dolbow and Belytschko (1999) propose an EFG formulation using selective reduced integration. Chen *et al.* (2000) suggest an improved RKPM based on a pressure projection method. These alternatives have the same advantages and inconveniences as in standard FEMs. Perhaps it is worth noting that in meshfree methods it is nontrivial to verify analytically the LBB condition for a given approximation of velocity and pressure.

One alternative that uses the inherent properties of meshfree methods and does not have a counterpart in finite elements is the pseudo-divergence-free approach (Vidal, Villon and Huerta, 2002; Huerta, Vidal and Villon, 2004b). This method is based on diffuse derivatives (see Section 2.2.6), which converge to the derivatives of the exact solution when the radius of the support, ρ , goes to zero (for a fixed ratio ρ/h). One of the key advantages of this approach is that the expressions of pseudo-divergence-free interpolation functions are computed a priori, that is, prior to determining the specific particle distribution. Thus, there is no extra computational cost because only the interpolating polynomials are modified compared with standard EFG.

4 RADIAL BASIS FUNCTIONS

Radial basis functions (RBF) have been studied in mathematics for the past 30 years and are closely related to meshfree approximations. There are two major differences between the current practice approaches in radial basis functions and those described in previous sections:

1. Most radial basis functions have noncompact support.
2. Completeness is provided by adding a global polynomial to the basis.

For these two reasons, the solution procedures usually have to be tailored for this class of global interpolants. Two techniques that avoid the drawbacks of global approximations are

1. Multipolar methods
2. Domain decomposition techniques

The most commonly used radial basis functions (with their names) are

$$\Phi_l(x) = \begin{cases} \|x - x_l\| = r & \text{linear} \\ r^2 \log r & \text{thin plate spline} \\ e^{-r^2/r^2} & \text{Gaussian} \\ (r^2 + R^2)^q & \text{multipolar} \end{cases}$$

where c , R , and q are shape parameters. The choice of these shape parameters has been studied by Kansa and Carlsson (1992), Carlsson and Foley (1991), and Rippa (1999).

Completeness of the radial basis function approximations is usually provided by adding a global polynomial to the approximation. For example, an approximation with quadratic completeness is

$$u(x) = c_0 + \sum_{j=1}^{n_d} c_j x_j + \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} d_{ij} x_i x_j + \sum_l a_l \Phi_l(x)$$

where c_0 , c_j , d_{ij} , and a_l are the unknown parameters.

One of the major applications of radial basis functions has been in data fitting. The reference by Carr, Fright and Beatson (1997), where millions of data points are fit, is illustrative of the power of this method. One of the first applications to the solution of PDEs is given by Kansa (1990), who used multiquadrics for smooth problems in fluid dynamics. In Sharan, Kansa and Gupta (1997), the method was applied to elliptic PDEs. In both cases, collocation was employed for the discretization. Exceptional accuracy was reported. Although a good understanding of this behavior is not yet available, evidently very smooth global approximants have intrinsic advantages over rough approximants for elliptic PDEs and other smooth problems (any locally supported approximant will have some roughness at the edge of its support). The low cost of RBF evaluation is another advantage. Wendland (1999) has studied Galerkin discretization of PDEs with radial basis functions. Compactly supported RBFs are also under development. Local error estimates for radial basis approximations of scattered data are given by Wu and Schaback (1993).

5 DISCONTINUITIES

One of the most attractive attributes of meshfree methods is their effectiveness in the treatment of discontinuities. This feature is particularly useful in solid mechanics in modelling cracks and shear bands.

The earliest methods for constructing discontinuous meshfree approximations were based on the visibility criterion (Organ *et al.*, 1996). In this method, in constructing the approximation at x , the nodes on the opposite side of the discontinuity are excluded from the index set S_x^v , that is, nodes on the opposite side of the

discontinuity do not affect the approximation at x . To be more precise, if the discontinuity is described implicitly (i.e. by a level set) by $f(x) = 0$, with $f(x) > 0$ on one side, $f(x) < 0$ on the other, then

$$\begin{cases} f(x_l) f(x) > 0 \Rightarrow l \in S_x^v \\ f(x_l) f(x) < 0 \Rightarrow l \notin S_x^v \end{cases}$$

The name 'visibility' criterion originates from the notion of considering a discontinuity as an opaque surface while choosing the nodes that influence the approximation at a point x , S_x^v . If a node x_l is invisible from x , then node l is not included in the index set S_x^v even when it falls within the domain of influence.

When a discontinuity ends within a domain, such as at a cracktip, the visibility criterion does not provide an adequate tool for constructing the discontinuity in the approximation. Around the cracktip the approximation must be constructed so that it is continuous in front of the cracktip but discontinuous behind it. One way to accomplish this is to include in the basis branch functions of the form

$$B_{li} = \begin{bmatrix} r^2 \sin \frac{\theta}{2}, r^2 \sin \frac{\theta}{2} \end{bmatrix} \quad (60)$$

where θ is the angle between the line to the cracktip and the tangent to the crack and r is the distance to the cracktip (see Figure 14(b)).

For cracks in elastic materials, the basis can be enriched by branch functions that span the asymptotic neartip field of the Westergaard solution (see Fleming *et al.*, 1997). The basis is then the polynomial basis plus the functions

$$B_{li} = \begin{bmatrix} \sqrt{r} \sin \frac{\theta}{2}, \sqrt{r} \sin \frac{\theta}{2} \sin \theta, \sqrt{r} \cos \frac{\theta}{2}, \\ \sqrt{r} \cos \frac{\theta}{2} \cos \theta \end{bmatrix} \quad (61)$$

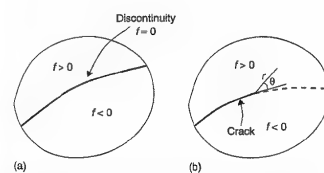


Figure 14. (a) Level set describing discontinuity and (b) nomenclature for cracktip branch functions.

Since this basis includes the first-order terms in the neartip field solution, very accurate solutions can be obtained with coarse discretizations. Xiao and Karahaloo (2003) have shown that even better accuracy can be attained by adding higher-order terms in the asymptotic field. The idea of incorporating special functions can also be found, among others, in Oden and Duarte (1997).

Regardless of whether a branch function is used near the cracktip, if the visibility criterion is used for constructing the approximation near the end of the discontinuity, additional discontinuities may occur around the tip. Two such discontinuities are shown in Figure 15. They can be avoided if the visibility criterion is not used near the tip. Kryst and Belytschko (1997) have shown that these discontinuities do not impair the convergence of the solution since their length, and hence the energy of the discontinuity, decreases with h . However, these extraneous discontinuities complicate the integration of the weak form, so several techniques have been developed to smooth the approximation around the tip of a discontinuity; the diffraction and the transparency method.

In the transparency method, the crack near the crack tip is made transparent. The degree of transparency is related to the distance from the crack tip to the point of intersection. The shape functions are smoothed around the crack tip as shown in Figure 16. The diffraction method is similar to the transparency method. The shape function is smoothed similar to the way light diffracts around a corner (see Organ *et al.*, 1996).

The recommended technique for including discontinuities is based on the partition of unity property of the approximants. It was first developed in the context of the extended finite element method (see Moes, Dolbow and Belytschko, 1999; Belytschko and Black, 1999; and Dolbow, Moes and Belytschko, 2000). It was applied in EFG by Ventura, Xu and Belytschko (2002). Let the set of particles whose support is intersected by the discontinuity be denoted by S^D , and the set of particles whose support includes the

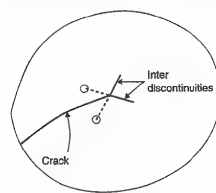


Figure 15. Inter discontinuity in visibility criterion method.

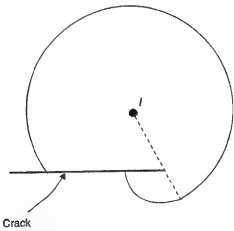


Figure 16. Domain of influence of node I by the transparency method near the crack tip.

tip of the discontinuity be S^T . The approximation is then given by

$$u^h(x) = \sum_{I \in S^T} N_I(x) u_I + \sum_{I \in S^0 \cap S^T} N_I(x) H(f(x)) q_I^{(1)} + \sum_{I \in S^0 \cap S^T} N_I(x) \sum_j (B_j \cdot q_j^{(2)})$$

where $H(\cdot)$ is the step function (Heaviside function). It is easy to show that the weak form (see Belytschko and Black, 1999) then yields the requisite strong form for elastic problems; this is also true for partial differential equations describing other physical problems.

The technique can also be used for intersecting and branching discontinuities (see Daux *et al.*, 2000; and Belytschko *et al.*, 2001). For example, consider the geometry shown in Figure 17. Let J^1 be the set of nodes whose domains of influence include the discontinuity $f_1(x) = 0$, J^2 the corresponding set for $f_2(x) = 0$. Let

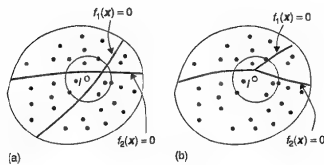


Figure 17. Support of node I with (a) intersecting discontinuities and (b) branching discontinuities.

$J^3 = J^1 \cap J^2$. Then the approximation is

$$u^h(x) = \sum_{I \in S^T} N_I(x) u_I + \sum_{I \in J^1 \cap S^T} N_I(x) H(f_1(x)) q_I^{(1)} + \sum_{I \in J^2 \cap S^T} N_I(x) H(f_2(x)) q_I^{(2)} + \sum_{I \in J^3 \cap S^T} N_I(x) H(f_1(x)) H(f_2(x)) q_I^{(3)}$$

5.1 Discontinuities in gradients

The continuously differentiable character of meshfree approximation functions is sometimes a disadvantage. At material interfaces in continua, or more generally, at discontinuities of the coefficients of a PDE, solutions of elliptic and parabolic systems have discontinuous gradients. These discontinuities in the gradient need to be explicitly incorporated in meshfree methods.

One of the first treatments of discontinuous gradients was by Cordes and Moran (1996). They treated the construction of the approximation separately; that is, they subdivided the domain into two subdomains. Let the subdomains Ω_1 and Ω_2 be with interface Γ^{int} . They enforced continuity of the function along Γ^{int} by Lagrange multipliers. Since the approximation $u(\Omega_1)$ was constructed without considering any nodes in Ω_2 and vice versa, the gradient of the approximation will be discontinuous across Γ^{int} after the continuity of the approximation is enforced by Lagrange multipliers. As in the case of essential boundary conditions, this continuity of the function can also be enforced by penalty methods, augmented Lagrangian methods or Nitsche's method.

An alternative technique was proposed by Krongauz and Belytschko (1998a), who added a function with a discontinuous gradient. In the original paper, the approximation is enriched with the absolute values of a signed distance function.

This enrichment function can also be employed with a local or global partition of unity

$$u(x) = \sum_I N_I u_I + N_I |f(x)| q_I$$

This was first studied in a finite element context by Sukumar *et al.* (2001) and Belytschko *et al.* (2001). In a finite element, the local partition of unity has some difficulties because the enrichment function does not decay, so it is difficult to fade it out gracefully. A global partition of unity is also undesirable. This behavior of this enrichment in meshfree methods has not been studied.

The local partition of unity method can also be used for intersecting gradient discontinuities, branching gradient discontinuities, and unclosed gradient discontinuities. The techniques are identical to those for discontinuities in functions, except that the step function is replaced by the absolute values of the signed distance function. For example, the approximation for a function with the intersecting discontinuities shown in Figure 17(a) is given by

$$u^h(x) = \sum_{I \in S^T} N_I(x) u_I + \sum_{I \in J^1 \cap S^T} N_I(x) |f_1(x)| q_I^{(1)} + \sum_{I \in J^2 \cap S^T} N_I(x) |f_2(x)| q_I^{(2)} + \sum_{I \in J^3 \cap S^T} N_I(x) |f_1(x)| |f_2(x)| q_I^{(3)}$$

The enrichment procedures are substantially simpler to implement than the domain subdivision/constraint technique of Cordes and Moran (1996), particularly for complex patterns such as intersecting and branching gradient discontinuities.

6 BLENDING MESHFREE METHODS AND FINITE ELEMENTS

Several authors have proposed different alternatives to blend finite elements and meshfree methods. These approximations are combined for two purposes: either *i)* to couple the two approximations, or *ii)* to enrich the finite element interpolation using particles.

In the first scenario, the objective is to benefit from the advantages of each interpolation in different regions of the computational domain, which is usually divided in three regions (see Figure 18). In one region, only finite elements are present, another with only meshfree approximation functions, and a transition region.

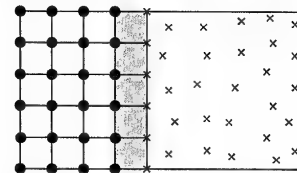


Figure 18. Discretization for the coupling of finite elements and a meshfree method: finite element nodes (\bullet), particles (\times) and transition region (in gray).

In the second case, the enrichment improves the approximation without remeshing. A finite element interpolation is considered in the whole domain and particles can be added in order to improve the interpolation in a selected region.

Coupled finite elements/meshfree methods are proposed by Belytschko, Organ and Krongauz (1995). They show how to couple finite elements near the Dirichlet boundaries and EFG in the interior of the computational domain. This simplifies considerably the prescription of essential boundary conditions. The mixed interpolation in the transition region requires the substitution of finite element nodes by particles and the definition of ramp functions. Thus the region for transition is of the size of one finite element (as in Figure 18) and the interpolation is linear. With the same objectives Hegen (1996) couples the finite element domain and the meshfree region with Lagrange multipliers.

Liu, Uras and Chen (1997b) independently suggest to enrich the finite element approximation with particle methods. In fact, the following adaptive process seems attractive: (1) compute an approximation with a coarse finite element mesh, (2) do an *a posteriori* error estimation and (3) improve the solution with particles without any remeshing process.

Following these ideas, Huerta and Fernández-Méndez (2000a) propose a unified and general formulation: the continuous blending method. This formulation allows both coupling and enrichment of finite elements with meshfree methods. The continuous blending method generalizes the previous ideas for any order of approximation, suppresses the ramp functions, and it does not require the substitution of nodes by particles. That is, as many particles as needed can be added where they are needed, independently of the adjacent finite element mesh. This is done in a hierarchical manner. This approach has been generalized in Chen *et al.* (2003) to get a nodal interpolation property.

Other alternatives are also possible; for instance, the bridging scale method proposed by Wagner and Liu (2000) is a general technique to mix a meshfree approximation with any other interpolation space, in particular with finite elements. Huerta, Fernández-Méndez and Liu (2004a) compare the continuous blending method and the bridging scale method, for the implementation of essential boundary conditions. The bridging scale method does not vanish between the nodes along the essential boundary. As noted by Wagner and Liu (2000), a modified weak form must be used to impose the essential boundary condition and avoid a decrease in the rate of convergence.

Next section is devoted to the continuous blending method proposed by Huerta and Fernández-Méndez (2000a). This method allows to recall the basic concepts on both enrichment and coupling. Although all the developments are done for the EFG method, the continuous

blending method is easily generalizable to other meshfree methods based on an MLS approximation.

6.1 Continuous blending method

Huerta and coworkers (Huerta and Fernández-Méndez, 2000a; Fernández-Méndez and Huerta, 2002; Fernández-Méndez, Díez and Huerta, 2003) propose a continuous blending of EFG and FEMs,

$$u(x) \simeq \tilde{u}(x) = \sum_{j \in \mathcal{J}} N_j^h(x) u_j + \sum_{i \in \mathcal{I}} \tilde{N}_i^f(x) u_i \quad (62)$$

$$= \pi^h u + \sum_{i \in \mathcal{I}} \tilde{N}_i^f(x) u_i$$

where the finite element shape functions $\{N_j^h\}_{j \in \mathcal{J}}$ are as usual, and the meshfree shape functions $\{\tilde{N}_i^f\}_{i \in \mathcal{I}}$ take care of the required consistency of the approximation, that is, consistency of order m . π^h denotes the projection operator onto the finite element space. Figure 19 presents two examples for the discretization. Particles are marked with \times and active nodes $\{x_i\}_{i \in \mathcal{I}}$ for the functional interpolation, are marked with \bullet . Other nonactive nodes are considered to define the support of the finite element shape functions (thus only associated to the geometrical interpolation). The first discretization (left) can be used for the coupling situation. Note that with the continuous blending method, the particles can be located where they are needed independently of the adjacent finite element mesh. In the second one (right), a finite element mesh is considered in the whole

domain and particles are added to enrich the interpolation, and increase the order of consistency, in the gray region in Figure 19.

The meshfree shape functions required in (62) are defined as in standard EFG (see equation (28))

$$\tilde{N}_i^f(x) = \alpha(x_i, x) \mathbf{P}^T(x_i) \tilde{\alpha}(x) \quad (63)$$

but now the unknown vector $\tilde{\alpha}$ is determined by imposing the reproducibility condition associated to the combined EFG and finite element approximation, that is,

$$\mathbf{P}(x) = \pi^h \mathbf{P}(x) + \sum_{i \in \mathcal{I}} \tilde{N}_i^f(x) \mathbf{P}(x_i) \quad (64)$$

Substitution of (63) in (64) leads to a small system of equations for $\tilde{\alpha} \in \mathbb{R}^{t+1}$ (see Huerta and Fernández-Méndez, 2000a for details)

$$\mathbf{M}(x) \tilde{\alpha}(x) = \mathbf{P}(x) - \pi^h \mathbf{P}(x) \quad (65)$$

The only difference with standard EFG is the modification of the r.h.s. of the previous system, in order to take into account the contribution of the finite element base in the approximation. Moreover, note that the expression for the modified EFG shape functions is independent of the situation, that is, the same implementation is valid for enrichment and coupling, increasing the versatility of this approach.

Figure 20 shows a 1D example of coupling finite elements and EFG. The meshfree shape functions adapt their shape to recover the linear interpolation. In the coupling

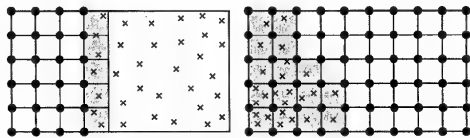


Figure 19. Possible discretizations for the continuous blending method.

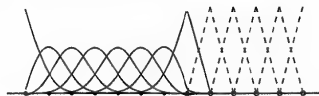


Figure 20. Shape functions of the coupled finite element (—) and meshfree (---) interpolation.

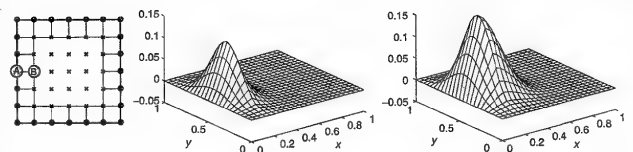


Figure 21. Discretization with active finite element nodes at the boundary (\bullet) and particles (\times), and meshfree shape function associated to the particle located at the gray circle (A) and (B) respectively.

situation, the continuity of the approximation is ensured under some conditions, even in multiple dimensions, by the following proposition (see Fernández-Méndez and Huerta, 2002, 2004 for the proof).

Proposition 3. The approximation $\tilde{u}(x)$, defined in (62), is continuous in Ω if (1) the same order of consistency m is imposed all over Ω (i.e. m coincides with the degree of the FE base), and (2) the domain of influence of particles coincides exactly with the region where finite elements do not have a complete basis.

Remark 24 (Imposing Dirichlet boundary conditions) Under the assumptions (1) and (2) in Proposition 3, the contribution of the particles is zero in the regions, where the finite element base is complete. In particular, this means that $\tilde{N}_i^f = 0$ in the finite element edges (or faces in 3D) whose nodes are all in \mathcal{J} (active nodes). This is an important property for the implementation of essential boundary conditions. If a finite element mesh with active nodes at the essential boundary is used, the meshfree shape functions take care of reproducing polynomials up to degree m in Ω and, at the same time, vanish at the essential boundary. For instance, Figure 21 shows the particle shape functions, $\tilde{N}_i^f = 0$, associated to particles at the boundary and in the first interior layer. Note that they vanish along the boundary because the finite element base is complete on $\partial\Omega$. Therefore, the prescribed values can be directly imposed as usual in the framework of finite elements, by setting the values of the corresponding nodal coefficients. Moreover, it is also easy to impose that the test functions (for the weak forms) vanish along the Dirichlet boundary (see Fernández-Méndez and Huerta, 2004 for further details).

The problem described in Figure 9 can be solved using the continuous blending method, Figure 22 shows the solution. As observed in Remark 24, the prescribed displacements are directly imposed. Two different finite element discretizations are considered. In both cases, the linear finite

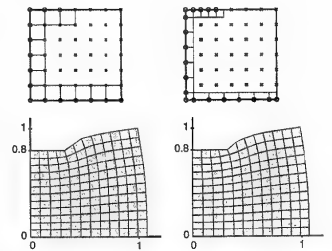


Figure 22. Continuous blending for two different distributions of finite elements near the essential boundary and the same distribution of particles, $h = 1/6$.

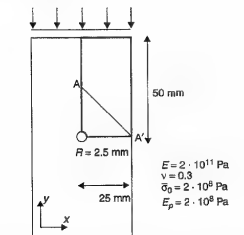


Figure 23. Problem statement: rectangular specimen with one centered imperfection.

element approximation at the boundary, allows the exact enforcement of the prescribed displacement. Note that if

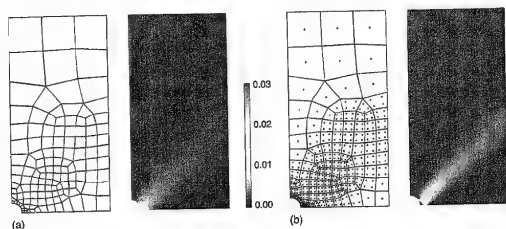


Figure 24. Coarse finite element mesh (Q1 elements) with its corresponding equivalent inelastic strain (b) and mixed interpolation with its equivalent inelastic strain distribution (a). A color version of this image is available at <http://www.mrw.interscience.wiley.com/cecm>

the prescribed displacement is piecewise linear or piecewise constant, as it is in this example, then it is imposed exactly when a bilinear finite element approximation is used.

Finally, an example reproduces the finite element enrichment with EFG in a nonlinear computational problem. A rectangular specimen with an imperfection is loaded (see Díez, Arroyo and Huerta, 2000; Huerta and Díez, 2000). Figure 23 presents the problem statement with the material properties.

This problem has been solved with the element-free Galerkin method. A coarse mesh solution of quadrilateral bilinear finite elements (308 dof) is shown in Figure 24 (left). When particles are added (308 + 906 = 1214 dof) and the order of consistency is increased ($m = 2$), an accurate distribution of inelastic strains is recovered (see Figure 24).

In this example, Figure 24, the original mesh is maintained and particles are added where they are needed.

REFERENCES

- Aluru N. A point collocation method based on reproducing kernel approximations. *Int. J. Numer. Methods Eng.* 2000; 47(6):1083–1121.
- Arnold DN, Brezzi F, Cockburn B and Marini LD. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.* 2001; 39(5):1749–1779.
- Askes H, Borst R and Heeres O. Conditions for locking-free elasto-plastic analyses in the Element-Free Galerkin method. *Comput. Methods Appl. Mech. Eng.* 1999; 173(1–2):99–109.
- Athuri S and Shen S. *The Meshless Local Petrov-Galerkin (MLPG) Method*. Tech Science Press: Encino, 2002.
- Babuska I. The finite element method with Lagrange multipliers. *Numer. Math.* 1973a; 20:179–192.
- Babuska I. The finite element method with penalty. *Math. Comp.* 1973b; 27:221–228.
- Babuska I and Melnik JM. *The Partition of the Unity Finite Element Method*. Technical Report BN-1185, Institute for Physical Science and Technology, University of Maryland, Maryland, 1995.
- Babuska I, Banerjee U and Osborn J. On principles for the selection of shape functions for the generalized finite element method. *Comput. Methods Appl. Mech. Eng.* 2002a; 191:5595–5629.
- Babuska I, Banerjee U and Osborn JE. Meshless and generalized finite element methods: a survey of some major results. In *Meshfree methods for partial differential equations*, vol. 26 of Lecture Notes in Computational Science and Engineering, Griebel M and Schweitzer MA (eds). Springer-Verlag: Berlin, 2002b; 1–20.
- Babuska I, Banerjee U and Osborn JE. Survey of meshless and generalized finite element methods: a unified approach. *Acta Numer.* 2003; 12:1–125.
- Babuska I, Caloz G and Osborn J. Special finite element methods for a class of second order elliptic problems with rough coefficients. *SIAM J. Numer. Anal.* 1994; 31:945–981.
- Becker R. Mesh adaptation for Dirichlet flow control via Nitsche's method. *Commun. Numer. Methods Eng.* 2002; 18(9):669–680.
- Belytschko T and Black T. Elastic crack growth in finite elements with minimal remeshing. *Int. J. Numer. Methods Eng.* 1999; 45(5):601–620.
- Belytschko T and Organ D. Element-free Galerkin methods for dynamic fracture in concrete. In *Computational Plasticity, Fundamentals and Applications*, Oñate E, Owen DRJ and Hilton E (eds). CIMNE: Barcelona, 1997; 304–321.
- Belytschko T, Liu WK and Moran B. *Nonlinear Finite Elements for Continua and Structures*. John Wiley & Sons: Chichester, 2000.
- Belytschko T, Lu YY and Gu L. Element free Galerkin methods. *Int. J. Numer. Methods Eng.* 1994; 37(2):229–256.
- Belytschko T, Organ D and Krongauz Y. A coupled finite element-element-free Galerkin method. *Comput. Mech.* 1995; 17(3):186–195.
- Belytschko T, Moes N, Usui S and Parimi C. Arbitrary discontinuities in finite elements. *Int. J. Numer. Methods Eng.* 2001; 50(4):993–1013.
- Belytschko T, Krongauz Y, Fleming M, Organ D and Liu WK. Smoothing and accelerated computations in the element free Galerkin method. *J. Comput. Appl. Math.* 1996a; 74(1–2):111–126.
- Belytschko T, Krongauz Y, Organ D, Fleming M and Krysl P. Meshless methods: an overview and recent developments. *Comput. Methods Appl. Mech. Eng.* 1996b; 139(1–4):3–47.
- Bonet J and Kulasegaram S. Correction and stabilization of smooth particle hydrodynamics methods with applications in metal forming simulations. *Int. J. Numer. Methods Eng.* 2000; 47(6):1189–1214.
- Bonet J and Lok TSL. Variational and momentum preservation aspects of smooth particle hydrodynamic formulations. *Comput. Methods Appl. Mech. Eng.* 1999; 180(1–2):97–115.
- Bouillard Ph and Suleau S. Element-free Galerkin method for Helmholtz problems: formulation and numerical assessment of the pollution effect. *Comput. Methods Appl. Mech. Eng.* 1998; 162(1–4):317–335.
- Britkopf P, Rasseigneux A and Villon P. Mesh-free operators for consistent field transfer in large deformation plasticity. In *Book of abstracts of the 2nd European Conference on Computational Mechanics: Solids, Structures and Coupled Problems in Engineering*, Cracow, 2001.
- Brezzi F. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Rev. Française Automat. Informat. Recherche Opérat. Sér. Rouge* 1974; 8(R-2):129–151.
- Carlsson RE and Foley TA. The parameter r^2 in multipadic interpolation. *Comput. Math. Appl.* 1991; 21:29–42.
- Carr JC, Frighi WR and Beaton RK. Surface interpolation with radial basis functions for medical imaging. *IEEE Trans. Med. Imag.* 1997; 16(1):96–107.
- Chen J-S, Han W, You Y and Meng X. A reproducing kernel method with nodal interpolation property. *Int. J. Numer. Methods Eng.* 2003; 56(7):955–960.
- Chen JS, Pan C, Wu CT and Liu WK. Reproducing kernel particle methods for large deformation analysis of non-linear. *Comput. Methods Appl. Mech. Eng.* 1996; 139(1–4):195–227.
- Chen JS, Yoon S, Wang H and Liu WK. An improved reproducing kernel particle method for nearly incompressible finite elasticity. *Comput. Methods Appl. Mech. Eng.* 2000; 181(1–3):117–145.
- Chu YA and Moran B. A computational model for nucleation of solid-solid phase transformations. *Model. Simul. Mater. Sci. Eng.* 1995; 3:455–471.
- Cordes LW and Moran B. Treatment of material discontinuity in the element-free Galerkin method. *Comput. Methods Appl. Mech. Eng.* 1996; 139(1–4):75–89.
- Daux C, Moes N, Dolbow J, Sukumar N and Belytschko T. Arbitrary branched and intersecting cracks with the extended finite element method. *Int. J. Numer. Methods Eng.* 2000; 48(12):1741–1760.
- De S and Bathe KJ. The method of finite spheres. *Comput. Mech.* 2000; 25(4):329–345.
- Díez P, Arroyo M and Huerta A. Adaptivity based on error estimation for viscoplastic softening materials. *Mech. Cohesive-Frict. Mater.* 2000; 5(2):87–112.
- Díez G. Moving-least-square-particles-hydrodynamics I: Consistency and stability. *Int. J. Numer. Methods Eng.* 1999; 44(8):1115–1155.
- Dolbow J and Belytschko T. Volumetric locking in the element free Galerkin method. *Int. J. Numer. Methods Eng.* 1999; 46(6):925–942.
- Dolbow J, Moes N and Belytschko T. Discontinuous enrichment in finite elements with a partition of unity method. *Finite Elem. Anal. Des.* 2000; 36(3–4):235–260.
- Donea J and Huerta A. *Finite Element Methods for Flow Problems*. John Wiley & Sons: Chichester, 2003.
- Duarte CA and Oden JT. An h - p adaptive method using clouds. *Comput. Methods Appl. Mech. Eng.* 1996a; 139(1–4):237–262.
- Duarte CA and Oden JT. H - p clouds – an h - p meshless method. *Numer. Methods Partial Differ. Equations* 1996b; 12(6):673–705.
- Dyka CT. *Addressing Tension Instability in SPH Methods*. Technical Report NRL/MR/6384, NRL, Washington, 1994.
- Fernández-Méndez S and Huerta A. Coupling finite elements and particles for adaptivity: an application to consistently stabilized convection-diffusion. In *Meshfree Methods for Partial Differential Equations*, vol. 26 of Lecture Notes in Computational Science and Engineering, Griebel M and Schweitzer MA (eds). Springer-Verlag: Berlin, 2002; 117–129.
- Fernández-Méndez S and Huerta A. Imposing essential boundary conditions in mesh-free methods. *Comput. Methods Appl. Mech. Eng.* 2004; 193(12–14):1257–1275.
- Fernández-Méndez S, Díez P and Huerta A. Convergence of finite elements enriched with meshless methods. *Numer. Math.* 2003; 96(1):43–59.
- Fleming M, Chu YA, Moran B and Belytschko T. Enriched element-free Galerkin methods for crack tip fields. *Int. J. Numer. Methods Eng.* 1997; 40(8):1483–1504.
- Freund J and Stenberg R. On weakly imposed boundary conditions for second order problems. In *Proceeding of the International Conference on Finite Elements in Fluids – New Trends and Applications*, Venezia, 1995.
- Gingold RA and Monaghan JJ. Smoothed particle hydrodynamics: theory and application to non-spherical stars. *Mon. Notices R. Astron. Soc.* 1977; 181:375–389.
- Girault V and Raviart P-A. *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*. Springer-Verlag, Berlin, 1986.
- Goss J and Liu WK. Admissible approximations for essential boundary conditions in the reproducing kernel particle method. *Comput. Mech.* 1996; 19(2):120–135.
- Gresho PM and Sani RL. *Incompressible Flow and the Finite Element Method*, Vol. 1: Advection Diffusion, Vol. 2: Isothermal Laminar Flow. John Wiley & Sons: Chichester, 2000.
- Griebel M and Schweitzer MA. A particle-partition of unity method for the solution of elliptic, parabolic and hyperbolic PDEs. *SIAM J. Sci. Comput.* 2000; 22(3):853–890.

- Glütter FC and Liu WK. Implementation of boundary conditions for meshless methods. *Comput. Methods Appl. Mech. Eng.* 1998; 163(1-4):205-230.
- Gunzburger MD. *Finite Element Methods for Viscous Incompressible Flows. A Guide to Theory, Practice, and Algorithms*. Academic Press: Boston, 1989.
- Hansbo P and Larson MG. Discontinuous Galerkin methods for incompressible and nearly incompressible elasticity by Nitsche's method. *Comput. Methods Appl. Mech. Eng.* 2002; 191(17-18):1895-1908.
- Hao S, Liu WK and Belytschko T. Moving particle finite element method with global smoothness. *Int. J. Numer. Methods Eng.* 2004; 59(7):1007-1020.
- Hegen D. Element free Galerkin methods in combination with finite element approaches. *Comput. Methods Appl. Mech. Eng.* 1996; 135(1-2):143-166.
- Huerta A and Díez P. Error estimation including pollution assessment for nonlinear finite element analysis. *Comput. Methods Appl. Mech. Eng.* 2000; 181(1-3):21-41.
- Huerta A and Fernández-Méndez S. Enrichment and coupling of the finite element and meshless methods. *Int. J. Numer. Methods Eng.* 2000a; 48(11):1615-1636.
- Huerta A and Fernández-Méndez S. Locking in the incompressible limit for the element free Galerkin method. *Int. J. Numer. Methods Eng.* 2001; 51(11):1361-1383.
- Huerta A, Fernández-Méndez S and Díez P. Enrichissement des interpolations d'éléments finis en utilisant des méthodes de particules. *ESAIM-Math. Model. Numer. Anal.* 2002; 36(6):1027-1042.
- Huerta A, Fernández-Méndez S and Liu WK. A comparison of two formulations to blend finite elements and mesh-free methods. *Comput. Methods Appl. Mech. Eng.* 2004a; 193(12-14):1105-1117.
- Huerta A, Vidal Y and Villon P. Pseudo-divergence-free element free Galerkin method for incompressible fluid flow. *Comput. Methods Appl. Mech. Eng.* 2004b; 193(12-14):1119-1136.
- Hughes TJR. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Dover Publications: New York, 2000. Corrected reprint of the 1987 original (Prentice Hall Inc., Englewood Cliffs, N.J.).
- Johnson GR and Beissel SR. Normalized smoothing functions for SPH impact computations. *Comput. Methods Appl. Mech. Eng.* 1996; 39(16):2725-2741.
- Kansa EJ. A scattered data approximation scheme with application to computational fluid-dynamics-I and II. *Comput. Math. Appl.* 1990; 19:127-161.
- Kansa EJ and Carlsson RE. Improved accuracy of multiquadric interpolation using variable shape parameters. *Comput. Math. Appl.* 1992; 24:88-120.
- Krongauz Y and Belytschko T. Consistent pseudo-derivatives in meshless methods. *Comput. Methods Appl. Mech. Eng.* 1998a; 146(1-4):371-386.
- Krongauz Y and Belytschko T. EFG approximation with discontinuous derivatives. *Int. J. Numer. Methods Eng.* 1998b; 41(7):1215-1233.
- Kryl P and Belytschko T. Element-free Galerkin method: convergence of the continuous and discontinuous shape functions. *Comput. Methods Appl. Mech. Eng.* 1997; 48(3-4):257-277.
- Lancaster GM. Surfaces generated by moving least squares methods. *Math. Comput.* 1981; 3(37):141-158.
- Li S and Liu WK. Meshfree and particle methods and their applications. *Appl. Mech. Rev.* 2002; 55(4):1-34.
- Liszka T and Orkisz J. The finite difference method at arbitrary irregular grids and its application in applied mechanics. *Comput. Struct.* 1980; 11:83-95.
- Liu GR. *Mesh Free Methods: Moving Beyond the Finite Element Method*. CRC Press: Boca Raton, 2002.
- Liu WK, Jun S, Li S, Adey J and Belytschko T. Reproducing kernel particle methods for structural dynamics. *Int. J. Numer. Methods Eng.* 1995a; 38(10):1655-1679.
- Liu WK, Jun S and Zhang YF. Reproducing kernel particle methods. *Int. J. Numer. Methods Fluids* 1995b; 20(8-9):1081-1106.
- Liu WK, Belytschko T and Oden JT (eds). Meshless methods. *Comput. Methods Appl. Mech. Eng.* 1996a; 139(1-4):1-440.
- Liu WK, Chen Y, Jun S, Chen JS, Belytschko T, Pan C, Uras RA and Chang CT. Overview and applications of the reproducing kernel particle methods. *Arch. Comput. Methods Eng.* 1996b; 3(1):3-80.
- Liu WK, Li S and Belytschko T. Moving least square reproducing kernel methods Part I: Methodology and convergence. *Comput. Methods Appl. Mech. Eng.* 1997a; 143(1-2):113-154.
- Liu WK, Uras RA and Chen Y. Enrichment of the finite element method with the reproducing kernel particle method. *J. Appl. Mech., ASME* 1997b; 64:861-870.
- Lucy L. A numerical approach to the testing of the fission hypothesis. *Astron. J.* 1977; 82:1013-1024.
- Melenk JM and Babuska I. The partition of unity finite element method: basic theory and applications. *Comput. Methods Appl. Mech. Eng.* 1996; 139(1-4):289-314.
- Moes N, Dolbow J and Belytschko T. A finite element method for crack growth without remeshing. *Int. J. Numer. Methods Eng.* 1999; 46(1):131-150.
- Monaghan JJ. Why particle methods work. *SIAM J. Sci. Stat. Comput.* 1982; 3(3):422-433.
- Monaghan JJ. An introduction to SPH. *Comput. Phys. Commun.* 1988; 48(1):89-96.
- Nayroles B, Touzot G and Villon P. Generating the finite element method: diffuse approximation and diffuse elements. *Comput. Mech.* 1992; 10(5):307-318.
- Nitsche J. Über eine Variation zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen die keinen Randbedingungen unterworfen sind. *Abh. Math. Se. Univ.* 1970; 36:9-15.
- Oden JT and Duarte CA. Clouds, cracks and fem's. In *Recent Developments in Computational and Applied Mechanics*, Reddy BD (ed.). A volume in honour of J.B. Martin. CIMNE: Barcelona, 1997; 302-321.
- Ölalt E and Idelsohn S. A mesh-free finite point method for advective-diffusive transport and fluid flow problems. *Comput. Mech.* 1998; 21(4-5):283-292.
- Organ D, Fleming M, Terry T and Belytschko T. Continuous meshless approximations for nonconvex bodies by diffraction and transparency. *Comput. Mech.* 1996; 18(3):225-235.
- Orkisz J. Meshless finite difference method. I Basic approach. In *Proceedings of the IACM-Fourth World Congress in Computational Mechanics*, 1998, CIMNE.
- Perrone N and Kao R. A general finite difference method for arbitrary meshes. *Comput. Struct.* 1975; 5:45-58.
- Pironneau O. *Finite Element Methods for Fluids*. John Wiley & Sons: Chichester, 1989.
- Quartapelle L. *Numerical Solution of the Incompressible Navier-Stokes Equations*, vol. 113 of International Series of Numerical Mathematics. Birkhäuser-Verlag: Basel, 1993.
- Quarteroni A and Valli A. *Numerical Approximation of Partial Differential Equations*, vol. 23 of Springer Series in Computational Mathematics. Springer-Verlag: Berlin, 1994.
- Randles PW and Libersky LD. Smoothed particle hydrodynamics: some recent improvements and applications. *Comput. Methods Appl. Mech. Eng.* 1996; 139(1-4):375-408.
- Randles PW and Libersky LD. Normalized SPH with stress points. *Int. J. Numer. Methods Eng.* 2000; 48(10):1445-1462.
- Rippa S. An algorithm for selecting a good value for the parameter c in radial basis function interpolation. *Adv. Comput. Mech.* 1999; 11:193-210.
- Schweitzer MA. *A Parallel Multilevel Partition of Unity Method for Elliptic Partial Differential Equations*. Springer. Lecture Notes in Computational Science and Engineering, Berlin; 29: 2003.
- Sharan M, Kansa EJ and Gupta S. Application of the multiquadric method for numerical solution of elliptic partial differential equations. *Appl. Math. Comput.* 1997; 84:275-302.
- Sukumar N, Chopp DL, Moes N and Belytschko T. Modeling holes and inclusions by level sets in the extended finite-element method. *Comput. Methods Appl. Mech. Eng.* 2001; 190(46-47):6183-6200.
- Sui M. Analytical and computational assessment of locking in the h_p finite element method. *Comput. Methods Appl. Mech. Eng.* 1996; 133(3-44):347-371.
- Sweezy JW, Hicks DL and Attaway SW. Smoothed particle hydrodynamics stability analysis. *J. Comput. Phys.* 1995; 116:123-134.
- Teman R. *Navier-Stokes Equations. Theory and Numerical Analysis*. AMS Chelsea Publishing: Providence, 2001. Corrected reprint of the 1984 edition [North Holland, Amsterdam, 1984].
- Ventura G, Xu JX and Belytschko T. A vector level set method and new discontinuity approximations for crack growth by EFG. *Int. J. Numer. Methods Eng.* 2002; 54(6):923-944.
- Vidal Y, Villon P and Huerta A. Locking in the incompressible limit: pseudo-divergence-free element-free Galerkin. *Rev. Eur. élém. Finis.* 2002; 11(7/8):869-892.
- Vila JP. On particle weighted methods and smooth particle hydrodynamics. *Math. Models Methods Appl. Sci.* 1999; 9(2):161-209.
- Villon P. *Contribution à l'optimisation*. Thèse présentée pour l'obtention du grade de docteur d'état, Université de Technologie de Compiègne, Compiègne, 1991.
- Wagner GJ and Liu WK. Application of essential boundary conditions in mesh-free methods: a corrected collocation method. *Int. J. Numer. Methods Eng.* 2000; 47(8):1367-1379.
- Wagner GJ and Liu WK. Hierarchical enrichment for bridging scales and mesh-free boundary conditions. *Int. J. Numer. Methods Eng.* 2001; 50(3):507-524.
- Wells GN, Borst R and Sluys LJ. A consistent geometrically nonlinear approach for delamination. *Int. J. Numer. Methods Eng.* 2002; 54(9):1333-1355.
- Wendland H. Meshless Galerkin Methods using radial basis functions. *Math. Comp.* 1999; 68(228):1521-1531.
- Wendland H. Local polynomial reproduction and moving least squares approximation. *IMA J. Numer. Anal.* 2001; 21:285-300.
- Wu ZM and Schaback R. Local error-estimates for radial basis function interpolation of scattered data. *IMA. J. Numer. Anal.* 1993; 13(1):15-27.
- Xiao QZ and Karhaloo BL. Direct evaluation of accurate coefficients of the linear elastic crack tip asymptotic field. *Fatigue Fract. Eng. M* 2003; 26(8):719-729.
- Zhu T and Atluri SN. A modified collocation method and a penalty formulation for enforcing the essential boundary conditions in the element free Galerkin method. *Comput. Mech.* 1998; 21(3):211-222.

Chapter 11

Discrete Element Methods

Nenad Bićanić

University of Glasgow, Glasgow, Scotland

| | |
|---|-----|
| 1 Introduction | 311 |
| 2 Basic Discrete Element Framework and Regularization of Nonsmooth Contact Conditions | 314 |
| 3 Characterization of Interacting Bodies and Contact Detection | 317 |
| 4 Imposition of Contact Constraints and Boundary Conditions | 321 |
| 5 Modeling of Block Deformability | 324 |
| 6 Transition Continuum/Discontinuum, Fragmentation in Discrete Element Methods | 329 |
| 7 Time Integration – Temporal Discretization, Energy Balance, and Discrete Element Implementation | 331 |
| 8 Associated Frameworks and Developments | 333 |
| References | 335 |
| Further Reading | 337 |

1 INTRODUCTION

Discrete element methods comprise a set of computational modeling techniques suitable for the simulation of dynamic behavior of a collection of multiple rigid or deformable bodies, particles or domains of arbitrary shape, subject to continuously varying contact constraints. Bodies collide with one another, new contacts are established, while old contacts may be released, giving rise to changes in the contact status and contact interaction forces, which in turn

influence the subsequent movement of bodies. Therefore, issues related to the nonsmoothness in space (separate bodies) and in time (jumps in velocities upon collisions) need to be considered, as well as the application of the interaction law (e.g. nonpenetrability, friction). Typically, a configuration of the collection of bodies changes continuously under the action of some external agency and as a result of the interaction law between bodies, leading to a steady state configuration at the state of rest, if a static equilibrium is reached. Bodies can be considered as rigid or deformable – if they are rigid, the interaction law between bodies in contact is the only constitutive law considered, whereas for deformable bodies, an appropriate homogenized continuum constitutive law (e.g. elasticity, plasticity, fracturing) needs to be accounted for as well.

Computational modeling of multibody contacts (contact detection and resolution) is clearly the dominant issue in discrete element methods, although the same issue appears in the nonlinear finite element analyses of contact problems (see Chapter 6, Volume 2). When the number of bodies or domains in contact is relatively small, it is possible a priori to define groups of nodes, segments, or surfaces, which belong to a possible contact set. These geometric attributes are then continuously checked against one another and the kinematic resolution can be treated in a very rigorous manner. Bodies that are possibly in contact may be internally discretized by finite elements (Figure 1), and their material behavior can essentially be of any complexity.

Discrete element methods are specifically geared for simulations involving a *large number of bodies* and emphasis lies on the change of contact locations and conditions that cannot be defined a priori and that need to be continuously updated as the solution progresses. Discrete element methods also represent powerful frameworks in which very simple interaction laws between individual particles

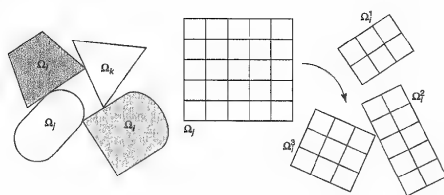


Figure 1. Collection of bodies, discretization of bodies into finite elements, changing configurations, and possible fragmentation.

can simulate the complex material behavior observed at a homogenized, macroscopic level.

The term *discrete element method* is most commonly associated with the definition (Cundall, 1989) that refers to any computational modeling framework, which

1. allows finite displacements and rotations of discrete bodies, including complete detachment and
2. recognizes new contacts automatically as the calculation progresses.

There exist a large number of methods or methodologies or procedures, which in one way or another belong to a broad class of discrete element methods. They are often referred to or could possibly be classified and distinguished according to the manner they deal with (a) contact detection algorithm, (b) treatment of contacts (rigid, deformable), (c) deformability and material model of bodies in contact (rigid, deformable, elastic, elasto-plastic etc.), (d) small-strain or large-strain formulations, (e) number (small or large) and distribution (loose or dense packing) of interacting bodies considered, (f) consideration of the model boundaries, (g) possible fracturing or fragmentation, and (h) time stepping integration schemes (explicit, implicit). Many methods differ in only a small number of the above attributes, yet they appear under different names. Moreover, the term discrete element methods is not used only in the case of preexisting discontinuities, as the *discrete* nature of the *emerging discontinuities* is often also taken into account.

Discontinuous modeling frameworks, which are increasingly utilized in modeling discontinuous, fractured and disjointed media also include techniques appearing under various names such as the rigid block spring method, discontinuous deformation analysis, combined discrete/finite elements, nonsmooth contact dynamics, and so on. Their applications (Figure 2) range from modeling problems of an inherently discontinuous behavior (granular

and particulate materials, silo flow, sediment transport, jointed rocks, stone or brick masonry) to problems in which the modeling of transition from a continuum to a discontinuum is more important. Increased complexity of different discontinuous models is achieved by incorporating the deformability of solid material and/or by more complex contact interaction laws and by the introduction of some failure or fracturing criteria controlling the solid material behavior and the emergence of new discontinuities.

On a homogenized continuum level, complex nonlinear continuum finite element analyses are conducted using inelastic material models and including, if appropriate, joint and interface elements to model any planes of weaknesses (see Chapter 10, Volume 2). Typically, if a small number of discontinuities needs to be considered, these interface elements are adopted within an overall nonlinear continuum modeling framework – on the other hand, if the number of discontinuities is very large, some form of homogenization is usually employed (see Chapter 12, Volume 2).

Most media can be treated as discontinuous at some level of observation (nano, micro, meso, macro), where the continuum assumptions cease to apply. This happens when the scale of the problem becomes similar to the characteristic length scale of the associated material structure and the interaction laws between bodies or particles are invoked instead of the continuum constitutive law. This chapter is concerned with the discontinuous modeling of interaction phenomena observed at a macro level, although similar arguments can be applied at various levels of observation.

Computational modeling of *macroscopically* particulate and inherently discontinuous media may not be dealt with in an adequate manner by a homogenized continuous description, and the discrete nature of discontinuities needs to be taken into account. Such analyses usually concern a system of multiple bodies or particles (rigid or deformable),

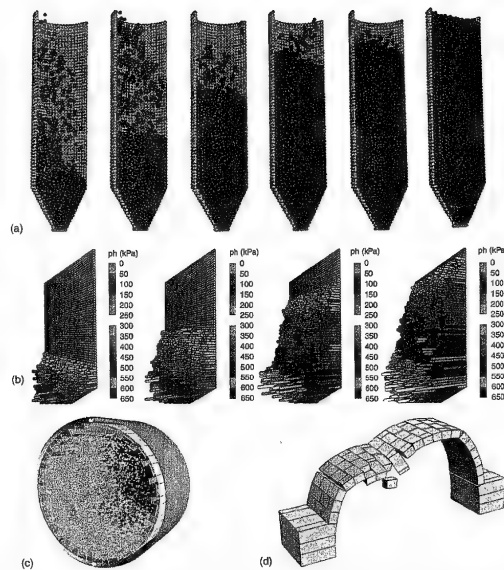


Figure 2. Typical discrete element applications – 3-D Hopper flow (a) and development of Silo wall pressures (b) during filling and discharge (after Lazarević D and Dvornik J. Selective time steps in predictor-corrector methods applied to discrete dynamic models of Granular materials. In ICADD-4, 4th Conference on Analysis of Discontinuous Deformation, Bičanić N (ed.), University of Glasgow, 2001), milling simulation (c) (after Cleary, 2002), 3-D masonry arches (d) (after Lemos JV. Assessment of the ultimate load of masonry arch using discrete elements. In *Comp Meth in Struct Masonry 3*, Middleton J and Pande G (eds), Books and Journals International, 1995). A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

which are potentially coming into contact as the solution progresses and the bodies to be considered can in general be of arbitrary shapes (e.g. granular materials). The solution evolves in time and is typically treated as a dynamic problem, which may or may not have a steady state solution. For discontinuous media, for example, jointed rocks or masonry, discontinuities are either preexisting (e.g. joints, bedding planes, interfaces, planes of weakness, construction joints) or emerging, in particular, in the case of cohesive frictional materials, where the growth and coalescence of micro-cracks eventually appear in a form of a

macro-crack. Many structures, structural systems, or structural components comprise discrete discontinuities, which need to be taken into account, where discontinuities may be heterogeneous or highly regular or structured. An obvious example of structured discontinuities is the brick masonry, or jointed rock structures in which the displacement discontinuities commonly occur at block interfaces, without necessarily rendering structures unsafe. Other structures exhibiting macro discontinuities (stone masonry, cracked structures, dilatation, or expansion joints) fall into a similar category.

This chapter will treat several important aspects of various discrete element methods such as:

- Body geometry characterization and contact detection
- Imposition of contact constraints and boundary conditions
- Definition and description of deformability
- Fracturing and fragmentation, transition from continua to discontinua
- Time stepping solution schemes

It will be shown that there are many similarities between the apparently different methods, and in the following, the most commonly encountered discrete element methods will be considered in the way they encompass the above aspects. Associated discontinuous modeling frameworks will also be discussed.

2 BASIC DISCRETE ELEMENT FRAMEWORK AND REGULARIZATION OF NONSMOOTH CONTACT CONDITIONS

In broader terms, the discrete element methods deal with either rigid or deformable particles. If the particles are rigid and are of a simple shape, an *event-by-event* simulation strategy can be applied. In such cases, collision times for particles can be calculated exactly and the *momentum exchange methodologies* are used to determine postcollision velocities, as the contact time is considered to be infinitely short. Any energy loss during contact is accounted for via the restitution coefficients or friction and the simulation deals with the nonsmooth step changes and reversals in velocities. Such methodologies have been used for molecular dynamics simulations with very large number of particles, and a range of contact detection and visualization techniques have been developed. However, although the event-driven algorithms work well for loose (gas-like)

assemblies of particles, for dense configurations these lead to an effective solution locking, that is, critically slow simulations, a phenomenon referred to as an *inelastic collapse* (McNamara and Young, 1994).

Collision of deformable bodies implies that the contact time is not infinitely short and that contact forces vary for the duration of the contact. Any simulation strategy therefore calls for some form of time stepping scheme and some way of *regularizing* the nonsmooth nature of the nonpenetration and friction condition.

Constraints of nonpenetration during the contact between the two bodies – termed here as a *contactor* Ω_c and a *target* body Ω_t – implies that no material point belonging to the contactor body should cross the boundary of the target body, that is, the gap between them must be nonnegative.

Only a compressive (here assumed positive) interaction force F_n is possible, that is, no attraction force between the two bodies exists and this interaction force vanishes for nonactive contact $g > 0$. These strict conditions of a unilateral contact can be mathematically described by the so-called Signorini condition. The above infinitely steep (i.e. 'nonsmooth') graph (Figure 3) can be regularized by assuming that the interaction force F_n is a function of the gap violation, which can be physically interpreted through elastic properties of an assumed contact layer. The infinitely

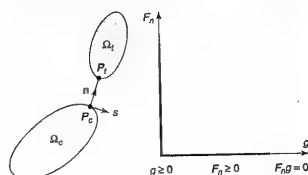


Figure 3. Nonsmooth treatment of normal contact.

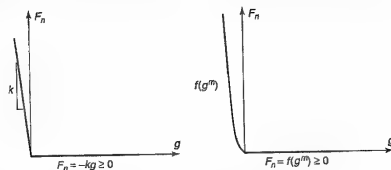


Figure 4. Regularized treatment of normal contact, with a linear and nonlinear penalty term.

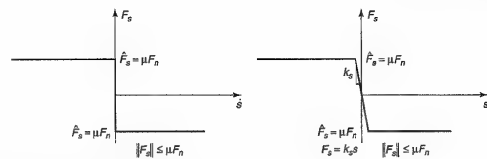


Figure 5. Nonsmooth and regularized treatment of frictional contact.

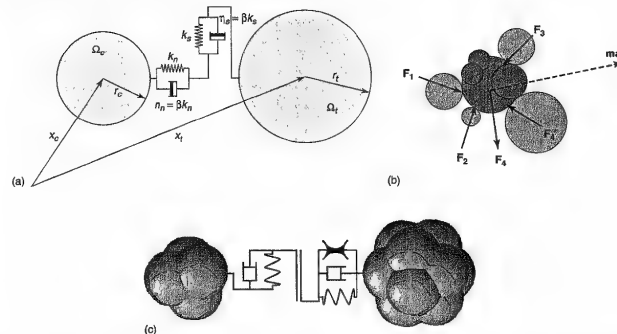


Figure 6. (a) Discrete element bodies (particles) in contact, giving rise to axial and tangential contact forces. Force magnitudes related to the relative normal and tangential velocity and to relative normal and tangential velocity at the contact point. (b) Arbitrary particle shapes as assemblies of clustered particles of simple shapes. (c) Contact of clustered particles, including liquid bridge forces to simulate wet particles (Groger T, Tuzun U and Heyes D. Shearing of wet particle systems discrete element simulations. In *1st International PFC Symposium*, Itasca, 2002). A color version of this image is available at <http://www.mrw.interscience.wiley.com/edcm>

steep graph is thereby replaced by the penalty formulation, with a linear or nonlinear penalty coefficient (Figure 4).

Nonsmooth relations also exist if the interaction law considers a tangential friction force F_s related and opposed to the sliding velocity \dot{s} . For the Coulomb friction law, there is a threshold tangential force proportional to the normal interaction force $F_t = \mu F_n$, before any sliding can occur, corresponding again to an infinitely steep graph (Figure 5). Usual regularization in the case of small tangential displacements (proportional to sliding velocity) is to formulate the friction law such that the friction force is proportional to the relative tangential displacement.

As a result of the above regularizations, the behavior of the collection of bodies is now governed by a set of

differential equations, which can be solved by some time stepping technique, where very small time steps are used in order to ensure accuracy in enforcing very stiff interaction conditions.

The initial formulation of the discrete element method, originally called *distinct element method* or DEM (Cundall, 1971), was based precisely on such regularization concepts and on the assumption of rigid elements and deformable contacts. Later extensions to include local deformation have permitted more rigorous treatment of both the contact conditions and energy preservation requirements. Over a period of time, a number of more sophisticated models for both the solid material as well as contacts have been formulated within the discrete element context.

The overall algorithmic framework for the DEM (with regularized contact constraints) is conceptually straightforward and it has remained more or less the same since the method was first introduced. Viscous contact forces (proportional to the relative velocities in the normal and tangential direction) are also included, adding even more to the regularized nature of the contact conditions. More complex interaction force fields can also be considered (Figure 6). The method basically considers each body in turn and at any given time determines all forces (external or contact) acting on it. Any out-of-balance force (or moment)

induces an acceleration (translational or rotational), which then determines the movement of that body during the next time step.

The simplest computational sequence for the DEM (most often formulated in the 'leap-frog' format, see Table 1) typically proceeds by solving the equations of motion of a given discrete element using an explicit time marching scheme, while updating contact force histories as a consequence of contacts between different discrete elements and/or resulting from contacts with model boundaries.

Table 1. Computational sequence for discrete element code (based on Cundall PA and Strack ODL. A discrete numerical model for granular assemblies. *Geotechnique* 1979; 29:47-65).

| | |
|--|--|
| | $e_n = \frac{y_1 - y_2}{D} = (\cos \alpha, \sin \alpha)$ $e_t = (\sin \alpha, -\cos \alpha)$ <p>Body center (x_1, x_2), (y_1, y_2)</p> <p>Translational velocity \dot{x}_1, \dot{y}_1</p> <p>Rate of rotation $\dot{\theta}_1, \dot{\theta}_2$</p> |
| <p>Time</p> <p>$n-1$ n $n+1$ $n+2$</p> <p>Δt</p> | |
| <p>FORCE-DISPLACEMENT LAW</p> | |
| <p>(1) Relative velocities</p> <p>(2) Relative displacements</p> <p>(3) Contact force increments</p> <p>(4) Total forces</p> <p>(5) Check for slip</p> <p>(6) Compute moments</p> | $\dot{X}_i = (\dot{x}_i - \dot{y}_i) - (\dot{\theta}_i R_x + \dot{\theta}_j R_y)t_i$ $\dot{h} = \dot{X}_i e_n, \dot{s} = \dot{X}_i e_t$ $\Delta n = \dot{h} \Delta t, \Delta s = \dot{s} \Delta t$ $\Delta F_n = k_n \Delta n + \beta k_n \dot{h}, \Delta F_t = k_t \Delta s + \beta k_t \dot{s}$ $F_n = F_n^{n-1} + \Delta F_n, F_t = F_t^{n-1} + \Delta F_t$ $F_s = \min(F_t, C + F_n \tan \phi)$ $M_x = \sum F_y R_x, M_y = \sum F_x R_y$ |
| <p>EQUATIONS OF MOTION</p> | |
| <p>(1) Assume force and moment constant over</p> <p>(2) Acceleration</p> <p>(3) Velocity</p> <p>(4) Assume velocities constant over</p> <p>(5) Displacements</p> <p>(6) Rotation</p> | $\Delta t (t^{n+1/2} - t^{n-1/2})$ $\ddot{x}_i^n = \frac{\sum F_x}{m}, \ddot{\theta}^n = \frac{\sum M_i}{I}$ $\dot{x}_i^{n+1/2} = \dot{x}_i^{n-1/2} + \ddot{x}_i^n \Delta t, \dot{\theta}^{n+1/2} = \dot{\theta}^{n-1/2} + \ddot{\theta}^n \Delta t$ $\Delta t (t^n - t^{n-1})$ $x_i^{n+1} = x_i^n + \dot{x}_i^{n+1/2} \Delta t$ $\theta^{n+1} = \theta^n + \dot{\theta}^{n+1/2} \Delta t$ |
| <p>TIME INCREMENT</p> | |

3 CHARACTERIZATION OF INTERACTING BODIES AND CONTACT DETECTION

Computational time step in the realization of the discrete element method requires a detection of bodies in contact and the evaluation of the contact forces (both magnitude and direction) emanating from the contact. If the interacting bodies are of very simple geometry (e.g. circular (2-D) or spherical (3-D)), these issues are straightforward, as the algorithmic check for a possible overlap is simple and the definition of the contact plane is unambiguous. Bodies of more complex shapes can also be conveniently approximated by forming convenient clusters (Figure 6b) of rigidly connected circular or spherical particles, while the contact detection and resolution remain the same as for single particles. However, very often the interacting bodies of arbitrary geometry need to be considered, and the algorithmic complexity of the contact detection and the associated definition of the contact plane between the two bodies increase significantly.

An efficient contact detection algorithm and a rational contact model to evaluate contact forces, for a large number of bodies where the relative position and shape of these arbitrary bodies may be continuously changing, is needed irrespective of the level of complexity adopted for the material description or discretization.

Within the concept of the discrete element formulation, each element is considered as a separate, distinct body,

which may or may not be in contact with various neighboring elements; hence, the main computational effort in such formulation is spent on the contact detection, that is, algorithms to establish which other bodies are in contact with the currently inspected body. The efficiency of these algorithms is crucial, as the conceptually simple procedure to test the possibility of contact of an element with all other elements at every time step becomes highly uneconomical once the number of elements becomes large. Contact search algorithms are typically based on so-called *body based search* or a *space based search*. In the former, only the space in the vicinity of the specified discrete element is searched (and the search repeated only after a number of time steps), whereas the latter implies a subdivision of the total searching space into a number of overlapping windows.

Contact detection problem between bodies of arbitrary geometries can be formally stated as finding a contact or overlap of a given contactor body with a number of bodies from a target set of N bodies in R^n space (Fig. 7). As a consequence of a desire to deal with arbitrary geometric shapes, most algorithms typically employ a *two-phase* strategy. Initially, all bodies may be approximated by simpler geometric representations, which encircle the actual body (*bounding volume, bounding box or bounding sphere*), and the list of possible contact pairs is established using an efficient *global neighbor or region search* algorithm.

This is then followed by a detailed *local contact resolution* phase, where the potential contact pairs are examined by considering the actual body geometries. This phase is

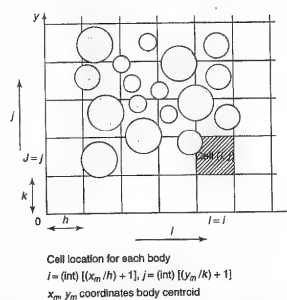


Figure 7. Basis of the hashing or binning algorithm for simple particle shapes and clustered particles. A color version of this image is available at <http://www.nrwl.interscience.wiley.com/cem>

strongly linked with the manner the geometry of actual bodies is characterized.

3.1 Global neighbor or region search

A typical example of the initial region search phase is the so-called boxing algorithm (Taylor and Preece, 1989). A complete computational domain is subdivided into regular intervals into cells, and a list of bodies overlapping a given cell (e.g. i, j in 2-D) is established via a contact detection of rectangular regions. Once this list is complete, the contact resolution phase for a given body comprises a detailed check against contact with all bodies listed to share the same cell and the check is usually extended to a list of bodies corresponding to a neighboring layer of cells (i.e. 8 cells in 2-D, 26 cells in 3-D).

A proper balance between the cell size and the maximum size of a body is clearly of the essence here. If the cell size is large compared to the body size, the initial search is fast, but many bodies may be listed as potential contact pairs and the contact resolution phase is likely to be extensive. On the other hand, if the cell size is small, the initial search is computationally demanding, which results in very few potential contact pairs and consequently, a less demanding contact resolution phase. A balance is reached with cell sizes that are approximately of the size of the largest body in the system.

Different options exist to formulate an efficient searching algorithm and the concepts are again frequently borrowed from related fields, typically comprising compact and efficient data representation techniques to describe the geometric position of the discrete element – for example, nodes, sides, or faces. The decomposition of the computational space and the efficiency of various cell data representation for a large number of contactor objects (binary tree, quad tree, direct evidence, combination of direct evidence, rooted trees, alternating data trees; Figure 8) are usually adopted (Taylor and Preece, 1989; Bonet and Peraire, 1991; Munjiza and Andrews, 1998; Petrinic, 1996; Williams and O'Connor, 1999; Perkins and Williams, 2001; Feng and Owen, 2002). Algorithmic issues and details of the associated data structures are quite involved (Williams and O'Connor, 1999), and the efficiency of many contact detection algorithms depend in a nonlinear fashion on the total number of bodies. Contact detection algorithms of linear complexity (i.e. linear dependence on the number of bodies) are desirable and indeed essential for simulations involving a very large number of bodies. A particularly detailed explanation, as well as the pseudocode for the so-called NBS (no binary search) algorithm for bodies of similar sizes (total contact detection time proportional to the total number of

bodies, irrespective of the particle packing density) can be found in (Munjiza and Andrews, 1998).

Very efficient data structures and representations can be used when simple geometries are considered, for example, when searching for a possible overlap of rectangles in 2-D or bounding boxes in 3-D. Typically, a given rectangular domain in R^n is characterized by a minimum set of parameters and then mapped into a representative point in an associated R^{2n} space (Figure 9). For example, a 1-D segment ($a-b$) may be mapped into a representative point in 2-D space, with coordinates (a, b) , or a 2-D rectangle of a size $(x_{\min} - x_{\max})$ and $(y_{\min} - y_{\max})$ may be mapped to a representative point in a 4-D space $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$. Alternative equivalent representation schemes are sometimes preferred, for example, characterizing a rectangular domain in R^2 by the starting point coordinates (x_{\min}, y_{\min}) and the two rectangle sizes (h_x, h_y) followed by a mapping into a different R^4 space $(x_{\min}, y_{\min}, h_x, h_y)$. As the representation of the physical domain is reduced to a point, region search algorithms can be more efficient in the representative R^{2n} spaces than in the physical R^n space. The easiest interpretation of the associated region search can be given for 1-D segments (Figure 9), but the concept can be generalized for 2-D and 3-D settings.

3.2 Contact resolution phase

Once the list of potential contact pairs is established, many different algorithms for the subsequent detailed contact resolution are possible. These algorithms depend greatly on the manner in which the bodies are characterized or described. The contact resolution is not only needed to confirm (or otherwise) whether the potential contact pair is indeed in contact – the contact resolution phase also establishes the orientation of the contact plane, so that a local (n, t, s) coordinate system can be determined and the conditions for impenetrability or sliding can be properly applied.

Typical geometry descriptors can be categorized into three main groups (Hogue, 1998) – (a) polygon or polyhedron representation, (b) implicit continuous function representation (elliptical or general superquadrics) and (c) discrete function representation (DFR).

If the geometry of polygonal domains in 2-D is defined in terms of corners and edges, a whole series of algorithms exist to determine an intersection of two coplanar polygons. Clearly, any restriction to convex polygons simplifies the algorithm considerably, as concave corners introduce additional complexities, with multiple contact points possible. However, in terms of defining the orientation of the contact plane there are no ambiguities in considering a corner-to-edge or an edge-to-edge contact, as the contact plane normal

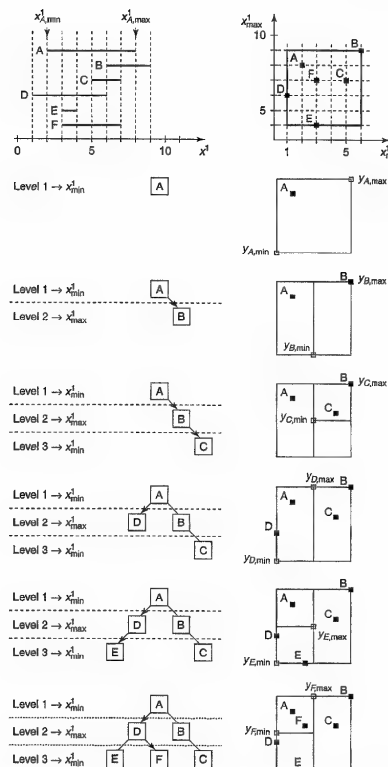


Figure 8. An example of the alternating data tree concept for storing object data (based on Petrinic N. *Aspects of Discrete Element Modelling Involving Facet-to-Facet Contact Detection and Interaction*, PhD thesis, University of Wales, Swansea, 1996).

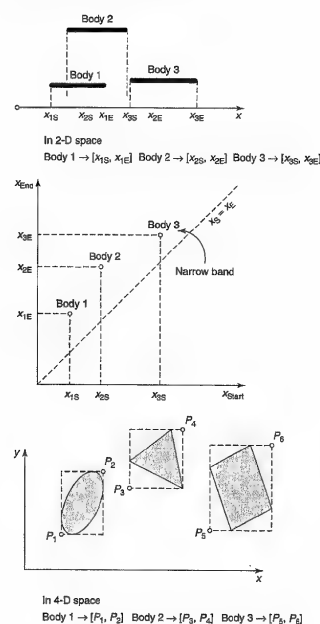


Figure 9. Mapping of a segment from 1-D space to a point in an associated 2-D space and mapping of a box in 2-D into 4-D space.

is obviously defined by the edge normal. Difficulties arise when the corner-to-corner contact needs to be resolved (Figure 10), as the orientation of the contact plane (and hence its normal) cannot be uniquely defined. This ambiguity can be avoided by the *rounding* of corners (Cundall, 1988) to ensure continuous changes of the contact outer normals, which was later enhanced through the introduction of a *common plane* (Cundall, 1988), 'hovering' between the two bodies that are coming into a corner-to-corner contact, whereby the actual orientation of this common plane is

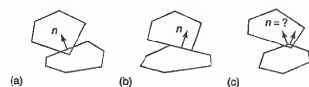


Figure 10. Definition of the contact plane – a unique definition for the corner-to-edge (a), edge-to-edge (b) case and an ambiguous situation for the corner-to-corner (c) contact problem (based on Hogue C. Shape representation and contact detection for discrete element simulations of arbitrary geometries. *Eng. Comput.* 1998; 3:374–390, copyright notice of Springer-Verlag).

found by maximizing the gap between the plane and a set of closest corners. Following that, a contact between a corner node and a plane is all that is needed for a robust contact resolution, with a vital benefit that the contact normal can always be determined.

Another possible procedure (restricted to 2-D situation) utilizes an optimum triangularization of the space between the polygons (Müller, 1996), whereby a collapse of a triangle indicates an occurrence of contact.

On the other hand, the *continuous implicit function representations* of bodies, for example, elliptical particles in 2-D (Ting, 1992; Vu-Quoc, Zhang and Walton, 2000), ellipsoids in 3-D (Lin and Ng, 1995), or superquadrics (Figure 11) in 2-D and 3-D (Pentland and Williams, 1989; Williams and Pentland, 1992; Wait, 2001)

$$\phi(x, y) = \left(\frac{x}{a}\right)^{p_1} + \left(\frac{y}{b}\right)^{p_2} - 1 \quad (1)$$

provide an opportunity to employ a simple analytical check (*inside-outside*) to identify whether a given point lies inside, or on the boundary $\phi(x, y) \leq 0$, or outside $\phi(x, y) > 0$ of the body. However, it is significantly more difficult to solve a complete intersection of contacting superquadrics, and the solution is normally found by discretizing one of the surfaces into facets and nodes, so the contact for a specific node can be verified through the inside-outside analytical check with respect to the functional representation of the other body.

A DFR utilizes the description of a body boundary via a parametric function of one parameter at distinct intervals. The concept of the DFR (O'Connor, Gill and Williams, 1993) arose essentially from the actual DEM implementation of the implicit function representation of bodies, where the calculation of the function values for the inside-outside check can be accelerated by preevaluating the function on a background grid, with scalar values assigned to each background grid node, which can then act as a fast algorithmic look-up table. As the discrete function values at the grid nodes need not necessarily stem from some implicit function, a grid (or cage) of cells can also be used to model

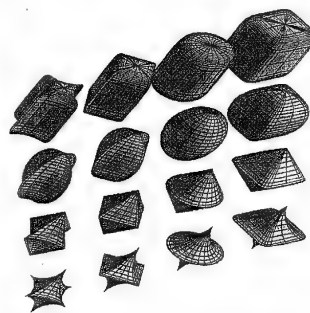


Figure 11. Superquadrics in 3-D (reproduced from Hogue C. Shape representation and contact detection for discrete element simulations of arbitrary geometries. *Eng. Comput.* 1998; 3:374–390, copyright notice of Springer-Verlag).

an arbitrarily shaped body – including bodies with holes (Figure 12).

Simplicity of the DFR concept is illustrated here (Figure 13) through the polar DFR descriptor in 2-D (Hogue and Newland, 1994), where, following the global

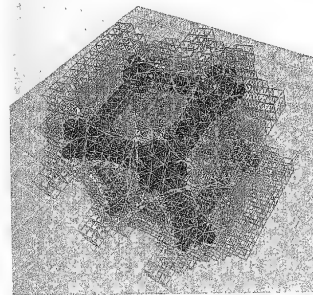


Figure 12. DFR representation of a 3-D object with holes (courtesy of Williams, MIT, IESL, and O'Connor (1999)). A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

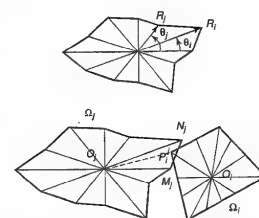


Figure 13. Contact detection in the polar discrete functional representation of bodies' geometry (after Hogue and Newland, 1994).

region search for possible neighbors, the local contact is established by transforming the local coordinates of the approaching corner P_i of a body i into the polar coordinates of the other body P_j and checking if no intersection between the segments $(O_i P_i)$ and $(M_j N_j)$ can be found.

4 IMPOSITION OF CONTACT CONSTRAINTS AND BOUNDARY CONDITIONS

4.1 Contact constraints between bodies

Once the contact between discrete elements is detected, the actual contact forces have to be evaluated, which in turn influence the subsequent motion of the discrete elements controlled by the dynamic equilibrium equations. Contact forces come about as a result of an imposition of contact constraints between the solution variables at contacting points. In variational formulations, constraint functional π_c can therefore be added to the functional of the unconstrained system in a variety of ways (see Chapter 6, Volume 2). The most frequently used *penalty format* includes a constraint functional $\pi_c = \int_{\Omega} \frac{1}{2} C^T(u) p C(u) d\Omega$, where $C(u) = Q$ is the constraint equation and p is the *penalty term*. No additional solution variables are required but the constraint equation is satisfied in an approximate sense, depending on the value of the penalty term. The use of large penalty terms clearly corresponds to a better imposition of the contact constraint, but such a choice corresponds to poorer conditioning of equations and has implications for the numerical integration schemes in terms of their stability and accuracy. On the other hand, the use of low penalty terms

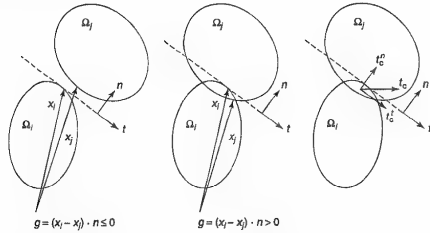


Figure 14. Determination of the contact surface and its local $n-t$ coordinate system.

leads to poorly satisfied contact constraints. Other forms of constraint functionals are possible, for example, the *Lagrangian multiplier method* $\pi_c = \int_{\Omega} \lambda^T C(u) d\Omega$, where the constraints are satisfied exactly, but an additional variable λ_i is introduced at every contact location. Modifications of the Lagrangian multiplier method come in a form of the *Perturbed Lagrangian method* $\pi_c = \int_{\Omega} \lambda^T C(u) d\Omega - \int_{\Omega} \lambda^T A \lambda d\Omega$ or most notably the *Augmented Lagrangian Method* $\pi_c = \int_{\Omega} \lambda^T C(u) d\Omega - \int_{\Omega} C^T(u) p C(u) d\Omega$, which combines the advantages of both the penalty and Lagrangian multiplier method, through an iterative update of Lagrange multipliers λ_i , without new variables introduced into the solution. The constrained Lagrangian approach is often adopted in quasi-static situations, but this appears to be far too expensive in the transient dynamic setting, as it requires an iterative sequence within every time step at every contact location.

Most discrete element formulation utilize the penalty function concept, which ultimately requires the information about the orientation of the contact surface (Figure 14) and its normal n as well as a geometric overlap or penetration of contactor objects to establish the orientation and the intensity of the contact forces between contactor objects at any given time, which in turn define the subsequent motion of the discrete elements. The nonpenetration condition is formulated through the gap function $g = [x^i - x^j] \cdot n \leq 0$, which leads to the relative displacement in the normal and tangential direction $u_n = (u_i - u_j) \cdot n$, $u_t = (u_i - u_j) \cdot t$ and the resolution of the total contact traction into $t_c = t_n \cdot n + t_t \cdot t = t_n^c + t_t^c$, which is then integrated over the contact surface to obtain the normal F_n and tangential component F_t of the contact force.

Whichever method is adopted, the imposition of the contact constraint is related to the normal and tangential directions associated with the orientation of the contact plane,

which is normally well defined, but clearly ambiguous in the case of the corner-to-corner contact. As no rigorous analytical solution exists, rounding of corners for arbitrary shaped bodies leads to an approximate Hertzian solution. In the case of a nonfrictional contact (i.e. normal contact force only), a robust resolution to the corner-to-corner problem in 2-D comes from the energy-based algorithm (Feng and Owen, 2002b), in which the concept of the contact energy potential is introduced. Contact energy W is assumed to be a function of the overlap area between the two bodies $W(A)$ and the contact force is oriented in the direction that corresponds to the highest rate of reduction of the overlap area A . As the overlap area is relative to bodies Ω_i and Ω_j , it can be expressed as a function of position of the corner point x_p and the rotational angle θ with respect to the starting reference frame.

Such an analytical process (Figure 15) leads to an unambiguous orientation of the contact plane in the 2-D corner-to-corner contact case, running through the intersection points g and h , and the contact force over the contact surface b_w needs to be applied through the *reference contact point* shifted by a distance $d = (M_\theta / \|F_n\|)$ from the corner, where F_n and M_θ are defined through the contact energy potential as $F_n = [\partial W(A) / \partial A][\partial A(x_p, \theta) / \partial x_p]$ and $M_\theta = [\partial W(A) / \partial A][\partial A(x_p, \theta) / \partial \theta]$. Different choices for the potential function are capable of reproducing various traditional models for contact forces.

| | $W(A)$ | $\ F_n\ $ |
|--------------------|-----------------------------|---------------------|
| Linear Form | $k_n A$ | $k_n b_w$ |
| Hertz Type Form | $\frac{2}{3} k_n A^{(3/2)}$ | $k_n A^{(1/2)} b_w$ |
| General Power Form | $\frac{1}{m} k_n A^m$ | $k_n A^{m-1} b_w$ |

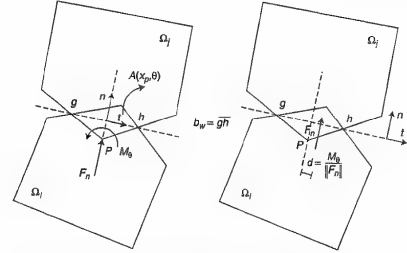


Figure 15. Corner-to-corner contact, based on energy potential (after Feng and Owen, 2002b).

The implementation of the nonlinear penalty term based on Hertz's analytical solution for circular deformable bodies has proved to be far superior (Thornton, 1992) to the adoption of a constant penalty term, especially in terms of avoiding spurious and artificial generation of energy during collisions, where the contact time within an increment is less than the computational time increment (see Chapter 6, Volume 2).

4.2 Contact constraints on model boundaries

An important aspect in the discrete element modeling is related to the representation and the treatment of model boundaries, for example, rigid or deformable, restrained or movable. Boundaries can be formulated either as real physical boundaries, or through virtual constraints. In the so-called *periodic boundary*, often adopted in the DEM

analysis of granular media, a virtual constraint implies that the particles (or bodies) 'exiting' on one side of the computational domain with a certain velocity are reintroduced on the other side, with the same, now 'incoming,' velocity. In cases of particle assemblies, the *flexible* (Kuhn, 1993) and *hydrostatic* boundaries (Ng, 2002) have been employed to improve the realism of a simulation as compared to the periodic boundary concept.

The flexible boundary framework (Figure 16) can be seen as a physical process of 'stringing' together particles on the perimeter of the particle assembly, forming a boundary network. Additional algorithmic issues arise related to an automatic identification of these perimeter particles and with updates of the boundary network as particles move. Flexible boundaries are mostly used to simulate the controlled stress boundary condition $\sigma_{ij} = \sigma_{ij}^c$, where the traction vector $t_i^m = A_m \sigma_{ij}^c n_j^m$ is distributed over the

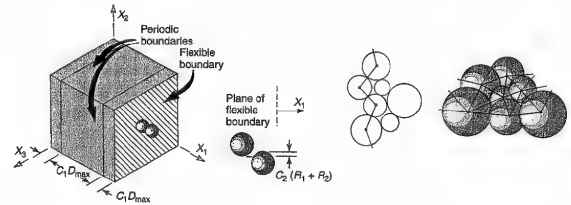


Figure 16. Flexible boundaries (reproduced from Kuhn M. A flexible boundary for three dimensional DEM particle assemblies. In 2nd International Conference on Discrete Element Methods, Williams J and Mustoe GWW (eds), MIT IESL, Publication, 1993).

centroids of three particles connected to form a triangular facet A_n .

The concept of the hydrostatic boundary is very simple (Figures 17 and 18), comprising a virtual wall of pressurized fluid imagined to surround granular material particles. If the particle shapes are characterized by some analytical expression (e.g. ellipsoid), intersection area with the virtual wall can easily be determined and the contact force is determined as $F_c = A_c p(h_c)$, where $p(h_c)$ is the prescribed

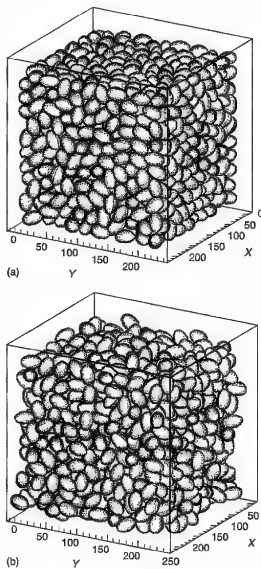


Figure 17. Examples of granular assemblies, comprising ellipsoidal particles, compacted under two idealized boundary conditions (a) hydrostatic and (b) periodic boundaries (reproduced from Ng TT. Hydrostatic boundaries in discrete element methods. In *Discrete Element Methods: Numerical Modeling of Discontinua*, 3rd International Conference on Discrete Element Methods, Cook BK and Jensen PJ (eds). ASCE Geotechnical Special Publication No. 117, 2002.)

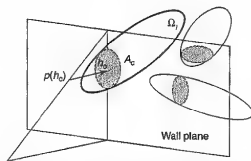


Figure 18. Concept of the hydrostatic boundary (after Ng TT. Hydrostatic boundaries in discrete element methods. In *Discrete Element Methods: Numerical Modeling of Discontinua*, 3rd International Conference on Discrete Element Methods, Cook BK and Jensen PJ (eds). ASCE Geotechnical Special Publication No. 117, 2002.)

hydrostatic pressure at the centroid of the intersection area A_c . Hydrostatic boundaries have also been used in combination with periodic boundaries.

It should be noted that the use of periodic boundaries excludes capturing of the localization phenomena and the introduction of real physical boundaries is required to account for these effects. Physical boundaries are also needed in problems in which boundary effects are important. Simplest form of boundary representation in 2-D is associated with a geometric definition of line segments (often referred to as *walls* in the DEM context), whereas the kinematics of the contact between the particle and the wall is again resolved in the penalty format. Frequently, individual particles are declared as nonmovable, thereby creating an efficient way of representing and characterizing a rigid boundary, without any changes in the contact detection algorithm. Recent successful ideas include the so-called finite wall (FW) method (Kremmer and Favier, 2001b) in which the boundary surface is triangulated into a number of rigid planar elements via a number of descriptor parameters, which are then in turn used to define an inscribed circle, as an efficient geometric object used in the contact detection analysis between the particles and the boundary.

Moreover, as the DEM analysis is usually set in a transient dynamics setting, problems with the treatment of artificial boundaries extending into infinite domains remain similar to the ones associated with the transient dynamic finite element analysis, that is, the issue of nonreflecting or transmitting boundaries needs to be taken into account (see Chapter 12, Volume 2).

5 MODELING OF BLOCK DEFORMABILITY

Consideration of increased complexities in the geometric characterization of rigid discrete element particles or bodies

(circles \rightarrow clusters of circles \rightarrow ellipses \rightarrow superquadrics \rightarrow general polygonal in 2-D; spheres \rightarrow clusters of spheres \rightarrow ellipsoids \rightarrow general superquadrics \rightarrow general polyhedra in 3-D) has made the method popular in many engineering applications. Subsequent DEM developments gradually introduced further complexities; in particular, the description of particle's deformability deserves attention.

Early attempts centered around superimposing a description of particle deformability on top of the rigid body movements, so that a displacement at any point within a simply deformable element can be expressed by $u_i = u_i^0 + \omega_{ij} x_j^0 + \epsilon_{ij} x_j^0$, where u_i^0 is the displacement of element centroid, ω_{ij} are the rotation and strain tensor respectively, and x_j^0 represent local coordinates of the point, relative to element centroid. Displacements of particle centroids emanate from standard equations for the translation of the center of mass in i th direction $\sum F_i = m \ddot{u}_i$, and the rotation about the center of mass $\sum M_i^c = I \ddot{\theta}_i$. The *simply deformable* discrete elements (Cundall, 1988) introduced equations for generalized strain modes ϵ_k , independent of the rigid body modes $m^k \ddot{e}^k = \sigma_k^a - \sigma_k^i$, where m^k is the generalized mass, σ_k^a is generalized applied stresses, and σ_k^i is generalized internal stresses, corresponding to strain modes. Different choices for the generalized strain field lead to different system matrices.

An alternative was suggested (Williams, Hocking and Mustoe, 1985) by introducing several body deformation mode shapes superimposed on top of the discrete body centroid motion. In that context, discrete element deformation field (displacement relative to the centroid) can also be expanded in terms of the eigenmodes of the generalized eigenvalue problem, associated with the discrete element stiffness and mass matrix, giving rise to the Modal Expansion Discrete Element Method (Williams and Mustoe, 1987). Here, the corresponding additional set of 'deformability' equations becomes decoupled (because of the orthogonality of eigenvectors) and modal amplitudes are solved for from simple scalar equations. The equations of motion are written with respect to the noninertial frame of reference, in order to ensure full decoupling.

Eventually, two predominant realizations of modeling block deformability appeared – the deformability of a discrete element of an arbitrary shape is either described by an internal division into finite elements (discrete finite elements or/and *combined finite/discrete elements*) or by a polynomial expansion of a given order (*discontinuous deformation analysis*).

5.1 Combined finite/discrete element method

An early combination of both discrete and finite elements was first successfully employed in the discrete

finite element approach (Ghaboussi, 1988). The combined finite/discrete element approach (Munjiza, Owen and Bićanić 1995) represents an extension of the rigid discrete element formulation, where the deformability is included via finite element shape-functions, that is, the problem is analyzed by the combination of the two methods. Such a combined method is particularly suited to problems in which progressive fracturing and fragmentation take place. In practice, the overall algorithmic framework for a combined finite element/discrete element framework (Table 2), remains largely realized in an explicit transient dynamic setting, that is, the scheme proceeds by solving the equations of motion using an explicit time marching scheme, while updating force histories as a consequence of contacts between different discrete regions, which are in turn subdivided into finite elements.

The deformability of individual discrete elements was initially dealt with by subdividing the body into triangular constant strain elements (Goodman, Taylor and Brekke,

Table 2. Simplified pseudocode for the combined discrete/finite element method, small displacement analysis, including material nonlinearity (after Petrinic N. *Aspects of Discrete Element Modelling Involving Facet-to-Facet Contact Detection and Interaction*, PhD thesis, University of Wales, Swansea, 1996).

- (1) Increment from the time station $t = t_n$
current displacement state u_n
external load vector, contact forces $F_n^{ext}, F_n^c \rightarrow \hat{F}_n^{ext}$
internal force, e.g. $F_n^{int} = \int_{\Omega} B^T \sigma_n d\Omega$
- (2) Solve for the displacement increment from
 $M \ddot{u}_n + F_n^{int} = \hat{F}_n^{ext}$
 $\ddot{u}_{n+1/2} = M^{-1} (\hat{F}_n^{ext} - F_n^{int}) \Delta t + \dot{u}_{n-1/2}$
 $u_{n+1} = u_n + \dot{u}_{n+1/2} \Delta t$
for an explicit time stepping scheme
 $m_i \ddot{u}_i^l + F_i^{int,l} = \hat{F}_i^{ext,l}$
 $\ddot{u}_{i+1/2}^l = \frac{1}{m_i} (\hat{F}_i^{ext,l} - F_i^{int,l}) \Delta t + \dot{u}_{i-1/2}^l$
 $u_{i+1}^l = u_i^l + \dot{u}_{i+1/2}^l \Delta t$
- (3) Compute the strain increment $\Delta \epsilon_{n+1} = f(\Delta u_{n+1})$
- (4) Check the total stress predictor $\sigma_{n+1}^* = \sigma_n + D \Delta \epsilon_{n+1}$ against a failure criterion, e.g. hardening plasticity $\phi(\sigma_{n+1}, \kappa) = 0$
- (5) Compute inelastic strain increment $\Delta \epsilon_{n+1}^{inel}$, e.g. associated plastic flow rule
- (6) Update stress state $\sigma_{n+1} = \sigma_n + D_{mod}(\Delta \epsilon_{n+1} - \Delta \epsilon_{n+1}^{inel})$
- (7) Establish contact states between discrete element domains at t_{n+1} and the associated contact forces F_{n+1}^c
- (8) $n \rightarrow n+1$, Go to step (1)

1968), which can be identified as an early precursor of a today's combined finite/discrete element modeling.

Large displacements and rotations of discrete domains, internally discretized by finite elements, have been rigorously considered (Barbosa and Ghaboussi, 1992) and typically the generalized Updated Lagrangian (UL) method is adopted. The contact forces, derived through the penalty formulation (concentrated and distributed contacts) and governed by simple constitutive relationship, are transformed into the equivalent nodal forces of the finite element mesh. The equations of motion for each of the deformable discrete elements (assuming also a presence of the mass proportional damping ($C = \alpha M$)) are then expressed as

$$M^e \ddot{U} + \alpha M^e \dot{U} = f_{\text{ext}} + f_{\text{cont}} - f_{\text{int}} = f_{\text{ext}} + f_{\text{cont}} - \sum_k \int_{\Omega_k} B_k^T \sigma_k d\Omega_k \quad (2)$$

Evaluation of the internal force vector $f_{\text{int}} = \sum_k \int_{\Omega_k} B_k^T \sigma_k d\Omega_k$ at the new time station $t + \Delta t$ recognizes the continuous changing of the configuration, as the Cauchy stress at $t + \Delta t$ cannot be evaluated by simply adding a stress increment due to straining of the material to the Cauchy stress at t , and the effects of the rigid body rotation on the components of the Cauchy stress tensor need to be accounted for.

In the computational realization, Barbosa and Ghaboussi (1992) used the central difference scheme for integrating the incremental updated Lagrangian formulation, which neglects the nonlinear part of the stress strain relationship. Incremental Green strain $\Delta \epsilon$ and the incremental 2nd Piola Kirchhoff stress $\Delta \Sigma$ are calculated from incremental displacements, which are then added to the 2nd Piola Kirchhoff stress, known from the old configuration, to be used subsequently to determine the internal force vector at the current configuration

$$f_{\text{int}}^{t+\Delta t} = \sum_k \int_{\Omega_k} B_k^T \Sigma_k d\Omega_k \quad (3)$$

In updating of the reference configuration, and in order to proceed to the next increment, the new deformation gradient F is required and the update of the Cauchy stress follows from

$$\sigma = \frac{1}{|F|} F^T \Sigma F \quad (4)$$

For inelastic analyses, due care needs to be given to the objectivity of the adopted constitutive law. Advanced combined discrete/finite element frameworks (Owen *et al.*,

Table 3. Simplified pseudocode for the combined discrete/finite element method, large displacement analysis (adapted from Petrinic, 1996).

- (1) Increment from the time station $t = t_n$
current displacement state u_n
external load vector, contact forces $F_n^{\text{ext}}, F_n^{\text{c}} \rightarrow \hat{F}_n^{\text{ext}}$
internal force, e.g. $F_n^{\text{int}} = \int_{\Omega} B^T \sigma_n d\Omega$
- (2) Solve for the displacement increment from
 $M \dot{u}_n + F_n^{\text{int}} = \hat{F}_n^{\text{ext}}$
 $\dot{u}_{n+1/2} = M^{-1}(\hat{F}_n^{\text{ext}} - F_n^{\text{int}}) \Delta t + \dot{u}_{n-1/2}$
 $u_{n+1} = u_n + \dot{u}_{n+1/2} \Delta t$
- (3) Configuration update $x_{n+1} = x_n + \Delta u_{n+1}$
- (4) Deformation Gradient $F_{n+1} = \frac{\partial x_{n+1}}{\partial x_n}$
- (5) Strain increment $\Delta \epsilon_{n+1} = f(F_{n+1}) = \frac{1}{2} F_{n+1}^T F_{n+1} - I$
- (6) Inelastic strain increment $\Delta \epsilon_{n+1}^{\text{inel}}$ from e.g.
 $\dot{\epsilon}_{n+1}^{\text{inel}} = \lambda \frac{\partial \Psi(\Sigma_{n+1})}{\partial \Sigma_{n+1}}$
- (7) Update 2nd Piola Kirchhoff stress state $\Sigma_{n+1} = \Sigma_n + D(\Delta \epsilon_{n+1} - \Delta \epsilon_{n+1}^{\text{inel}})$
- (8) Rotate stress state to obtain Cauchy stress $\sigma_{n+1} = R_{n+1} \Sigma_{n+1} R_{n+1}^T$
- (9) Establish contact states between discrete element domains at t_{n+1} and the associated contact forces F_{n+1}^{c}
- (10) $n \rightarrow n + 1$, Go to step (1)

1999) included a rigorous treatment of changes in configuration and evaluation of the deformation gradient and the objective stress measures; see Table 3.

5.2 Discontinuous deformation analysis, DDA

An alternative deformability representation is employed in the discontinuous deformation analysis (DDA), in which a general polynomial approximation of the strain field is superimposed to the centroid movement for each discrete body (Shi, 1988). The DDA analysis appeared initially as an efficient framework of modeling jointed deformable rock and its development followed the formulation of the Keyblock Theory (Shi and Goodman, 1981), which represents a procedure to assess stability limit states of rigid jointed rock blocks in 3-D. Blocks of arbitrary shapes with convex or concave boundaries, including holes are considered. The original framework under the name the 'DDA method' comprises a number of distinct features – (a) the assumption of the order of the displacement approximation field over the whole block domain, (b) the derivation of the incremental equilibrium equations on the basis of the minimization of potential energy, (c) the block interface

constitutive law (Mohr–Coulomb) with tension cutoff, and (d) use of a special implicit time stepping algorithm. The original implementation has been since expanded and modified by many other researchers, but the essential structure of the DDA framework has not substantially changed. The method is realized in an incremental form and it deals with the large displacements and deformations as an accumulation of small displacements and deformations. The issue of inaccuracies when large rotations are occurring has been recognized and several partial remedies have been proposed (McLaughlin and Sitar, 1996; Kc, 1996).

Leaving aside specific algorithmic features that constitute the DDA methodology, the method can best be seen as an alternative way of introducing solid deformability into the discrete element framework, where block sliding and separation are considered along predetermined discontinuity planes. Early formulation was restricted to simply deformable blocks (constant strain state over the entire block of arbitrary shape in 2-D, Figure 19), where the first-order polynomial displacement field for the block $[u \ v]^T$ can be shown to be equivalent to the three displacement components of the block centroid, augmented by the displacement field, which has a distinct physical meaning, which corresponds to the three constant strain states. All six variables are denoted by the block deformation vector D_{1st}^T .

$$\begin{bmatrix} u \\ v \end{bmatrix}_{\text{1st}}^T = \begin{bmatrix} 1 & 0 & -(y - y_0) & (x - x_0) & 0 & \frac{(y - y_0)}{2} \\ 0 & 1 & (x - x_0) & 0 & (y - y_0) & \frac{(x - x_0)}{2} \end{bmatrix}_I D_{\text{1st}}^T$$

$$[D_{\text{1st}}^T]^T = [u_0 \ v_0 \ \phi_0 \ \epsilon_x \ \epsilon_y \ \gamma_{xy}]^T \quad (5)$$

$$\begin{bmatrix} u \\ v \end{bmatrix}_{\text{1st}}^T = T_{\text{1st}}^T D_{\text{1st}}^T$$

Improved model deformability is achieved by either increasing the number of block deformation variables (higher-order DDA, where higher order strain fields are assumed for blocks of arbitrary shapes) or by the so-called subblock concept (Lin, 1995), in which a block is subdivided into a set of simply deformable subblocks.

In that spirit, the second-order approximation for the block displacement field requires 12 deformation variables, which can also be given a recognizable physical meaning – the deformation parameters comprise the centroid displacements and rotation, strain tensor components at the centroid,

$$D^T = [u \ v \ \phi \ \epsilon_x \ \epsilon_y \ \gamma_{xy}]$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1 & 0 & x & 0 & y & 0 & x^2 & 0 & xy & 0 & y^2 & 0 \\ 0 & 1 & 0 & x & 0 & y & 0 & x^2 & 0 & xy & 0 & y^2 \end{bmatrix} \begin{bmatrix} u_0 \\ v_0 \\ \phi_0 \\ \epsilon_x \\ \epsilon_y \\ \gamma_{xy} \\ \epsilon_{xx} \\ \epsilon_{yy} \\ \epsilon_{xy} \end{bmatrix} = T D$$

Figure 19. Deformation variables for the first- and second-order polynomial approximation in discontinuous deformation analysis. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecom>

as well as the spatial gradients of the strain tensor components, that is,

$$[D^T]_{\text{2nd}}^T = [u_0 \ v_0 \ \phi_0 \ \epsilon_x^0 \ \epsilon_y^0 \ \gamma_{xy}^0 \ \epsilon_{xx,x} \ \epsilon_{xx,y} \ \epsilon_{xy,x} \ \epsilon_{xy,y} \ \epsilon_{yy,x} \ \epsilon_{yy,y}]^T \quad (6)$$

For the higher-order approximations (Ma, Zaman and Zhou, 1996) for the block displacement field, it is difficult to give a clear physical interpretation to the deformation variables, and the generalized deformation parameters are adopted

$$u = d_1 + d_2 x + d_3 y + d_4 x^2 + d_5 xy + d_{11} y^2 + d_{x-1} x^n + d_{m-1} y^n$$

$$v = d_2 + d_4 x + d_6 y + d_8 x^2 + d_{10} xy + d_{12} y^2 + d_{x^n} x^n + d_{m^n} y^n$$

$$\begin{bmatrix} u \\ v \end{bmatrix}_{\text{nth}}^T = T_{\text{nth}}^T D_{\text{nth}}^T$$

$$[D_{\text{nth}}^T]^T = [d_1 \ d_2 \ d_3 \ \dots \ d_{m-1} \ d_m]^T \quad (7)$$

Once the block displacement field is approximated with a finite number of generalized deformation variables, the associated block strain and block stress field can be expressed in a similar manner as in the context of finite elements as

$$[\epsilon^T] = [B^T][D^T]$$

$$[\sigma^T] = [E^T][\epsilon^T] = [E^T][B^T][D^T] \quad (8)$$

For a system of N blocks, all block deformation variables (n variables per block, depending on the order of the

approximation) are assembled into a set of system deformation variables ($N \times n$) and the simultaneous equilibrium equations are derived from the minimization of the total potential energy. The total potential energy π comprises contributions from the block strain energy π_s , energy from the external concentrated and distributed loads π_p , π_q , interblock contact energy π_c , block initial stress energy π_0 , as well as the energy associated with an imposition of displacement boundary conditions π_b .

$$\pi = \pi_s + \pi_p + \pi_q + \pi_c + \pi_0 + \pi_b \quad (9)$$

Components of the stiffness matrix and the load vector are obtained by the usual process of the minimization of the potential energy

$$[K^{ij}] = \int_{\Omega} [B^i]^T [E^i] [B^j] d\Omega \quad (10)$$

The global system stiffness matrix contains ($n \times n$) submatrices K_{ij} and K_{ji} where the nonzero submatrices K_{ij} are present only if and when the blocks i and j are in active contact (Figure 20) and D comprises deformation variables of all blocks considered in the system.

The interblock contacts conditions of nonpenetration and Mohr–Coulomb friction can be interpreted as block displacement constraints, which is algorithmically reduced to

$$\begin{bmatrix} K_{11} & K_{12} & \dots & M_{1n} \\ K_{21} & K_{22} & \dots & K_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ K_{m1} & K_{m2} & \dots & m_{mn} \end{bmatrix} \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_n \end{bmatrix}$$

Figure 20. Assembly process in DDA analysis. A color version of this image is available at <http://www.mrw.interscience.wiley.com/cecm>

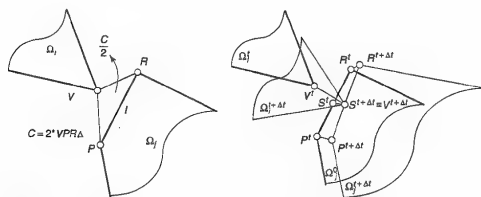


Figure 21. Nonpenetration and frictional contact constraint in DDA, point-to-edge contact.

an interaction problem between a vertex of one block and the edge of another block. If the deformation increments of the two blocks are denoted by D^i and D^j respectively, the nonpenetration of the vertex in the direction normal to the block edge can be expressed as a function of these deformation increments by

$$d = C + [D^i]^T [H^i] + [D^j]^T [G^j] = 0 \quad (11)$$

which represents the contact constraint for the block deformation variables D^i and D^j and where C is a function of the location of the vertex and the two end point of the block edge at the beginning of an increment. Various algorithmic approaches can be identified depending on which format is used for an implicit imposition of the nonpenetration condition (Figure 21).

If the penalty format is adopted, additional terms appear both in the global DDA stiffness matrix as well as on the

Table 4. Additional terms in the stiffness matrix and the load vectors as a result of contact between bodies i and j .

| Additional DDA stiffness matrix terms and changes in the load vector | | | |
|--|--|---------------------------------|--|
| Normal nonpenetration constraint | | | |
| $K^{ii} = K^{ii} + p[H^i][H^i]^T$ | | $F^i = F^i - pC[H^i]$ | |
| $K^{ij} = p[G^j][H^i]^T$ | | $F^j = F^j - pC[G^j]$ | |
| $K^{ji} = p[G^i][H^j]^T$ | | | |
| $K^{jj} = K^{jj} + p[G^j][G^j]^T$ | | | |
| Frictional constraint | | | |
| $K^{ii} = K^{ii} + p[H_{fr}^i][H_{fr}^i]^T$ | | $F^i = F^i - pC_{fr}[H_{fr}^i]$ | |
| $K^{ij} = p[H_{fr}^j][G_{fr}^i]^T$ | | $F^j = F^j - pC_{fr}[G_{fr}^i]$ | |
| $K^{ji} = p[G_{fr}^j][H_{fr}^i]^T$ | | | |
| $K^{jj} = K^{jj} + p[G_{fr}^j][G_{fr}^j]^T$ | | | |

RHS load vector (Table 4), and the terms differ depending on the nature of the constraint (normal nonpenetration or frictional constraint).

In both cases, the penalty formulation leads to a nonlinear iterative scheme that proceeds until the global equilibrium is satisfied (norm of the out-of-balance forces within some tolerance) while at the same time a *near zero* penetration condition is satisfied at all active contact positions. In the case of a normal nonpenetration condition, the convergence implies that the identified set of contacts does not change between iterations, whereas in the case of a frictional constraint, it implies that the changes in the location of the projected contact point remain within a given tolerance. For complex block shapes, the convergence process may sometimes be very slow indeed, as both activation and deactivation of contacts during the iteration process are possible. The convergence of the solution algorithm depends highly on the choice of the penalty term, and the process may often lead to ill-conditioned matrices if a very large penalty term is employed to ensure that the penetrations remain close to zero.

Alternatively, if the Lagrange multiplier method is used, the utilization of the constraint equation between blocks i and j and the solution requires the use of a special matrix pseudoinversion procedures. In the context of the Augmented Lagrange Multiplier Method, an iterative combination of a Lagrange multiplier and a contact penalty spring is utilized and the iteration proceeds until the penetration distance and the norm of the out-of-balance forces is not smaller than some specified norm.

6 TRANSITION CONTINUUM/DISCONTINUUM, FRAGMENTATION IN DISCRETE ELEMENT METHODS

Inclusion of fracturing and fragmentation to the discrete element method started in mid- and late-'80s (Mustoe, Williams and Hocking, 1987; Hocking, Mustoe and Williams, 1987; Williams, Mustoe and Hocking, 1987; Hocking 1989), including interelement and through-element brittle fracturing. Complexities associated with modeling of conditions for gradual fracturing (strain localization and strain softening prior to the eventual separation by cracking or shear slip) are the same for both the nonlinear finite element frameworks as well as for the combined finite element/discrete element simulations (De Borst, 2001) (see Chapter 10, Volume 2). The continuum-based material models for fracturing media are usually generalizations of elastoplasticity laws using different failure (or fracture) surface descriptions, and no discontinuities are admitted

in the displacement field as the geometry of a problem remains unchanged. Some models attempt to also simulate postfracturing behavior, again via continuum formulations. Models adopted for prefracturing, that is at a continuum stage, are usually based on concepts of damage mechanics, strain softening plasticity formulations (often utilizing fracture mechanics concepts of energy release required to open a crack or induce a shear slip), or have been formulated using some higher-order continuum theory.

Fracturing in DEM was typically confined to element interfaces, where models were either based on the fracture energy release rate concept or on the breakage of cohesive links between discrete elements. The combined finite/discrete element method (Munjiza, Owen and Bićanić, 1995) considers fracturing media, starting from a continuum representation by finite elements of the solid domain of interest, allowing for progressive fracturing to take place according to some fracturing criterion, thereby forming discontinuities and leading eventually to discrete elements that may be composed of several deformable finite elements. Subsequent motion of these elements and further fracturing of both the remaining continuum domain and previously created discrete elements are then modeled and monitored. In case of fracturing and fragmenting media, the main issues which require consideration are (a) finite element formulation capable of capturing the strain localization leading onto subsequent fracturing of the original continuum, (b) fracturing criteria and models, (c) remeshing algorithms for fully fractured zones, (d) contact detection procedures, and (e) representation of frictional contact conditions.

Improved formulation for nonlinear physical models and associated computational issues for DEM are closely related to advances in continuum-based computational plasticity, usually adopted in the FEM context. On the algorithmic front, improvements include more efficient contact detection and interaction algorithms, as well as the introduction of advanced computational concepts (parallel and distributed computing, object oriented programming, in core databases). Approaches for coupling discrete element methods with fluid flow methods have also appeared and a number of combined finite/discrete element simulations of various engineering problems have been reported, where the 'traditional' nonlinear explicit transient dynamic finite element and combined finite/discrete element formulations differ mainly in the number of contacts considered, the automatic creation of new discrete elements and the detection of new contacts.

An additional algorithmic problem arises upon separation, as the 'book keeping' of neighbors and updating of the discrete element list are needed whenever a new (partial or complete) failure occurs. In addition, there is also a

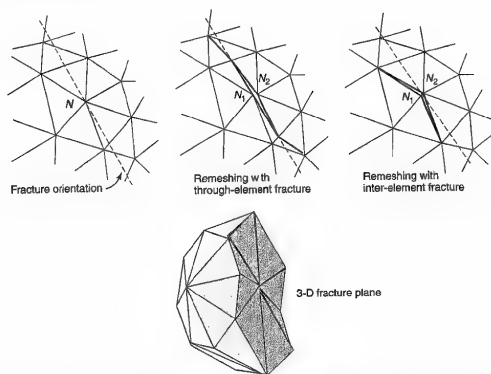


Figure 22. Element- and node-based local remeshing algorithm in 2-D and 3-D context (after Munjiza A, Owen DRJ and Bićanić N. A combined finite/discrete element method in transient dynamics of fracturing solids. *Eng. Comput.* 1995; 12:145–174; Owen *et al.*, 1999), based on a weighted local residual strength concept. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

need to transfer state variables (plastic strains, equivalent plastic strain, dissipated energy, damage variables) from the original deformable discrete element to the newly created deformable discrete elements.

Advances in continuum-based computational plasticity in terms of the consistent linearization required to secure a quadratic convergence of the nonlinear solution procedure have also influenced algorithms used in the DEM context. More accurate and robust stress return algorithms for softening plasticity models are employed, involving complex smooth surface descriptions, as well as surfaces with singular regions, where the computational features are frequently borrowed from various nonlinear optimization algorithms.

After fragmentation, every time a partial fracture takes place, the discrete element changes its geometry, and a complete fracture leads to the creation of two or more discrete elements, there is a need for automatic remeshing of the newly obtained domains (Figure 22). An unstructured mesh regeneration technique is preferred in such cases, where the mesh orientation and mesh density can sometimes be decided upon on the basis of the distribution of the residual material strength within the damaged rock or some other state variable.

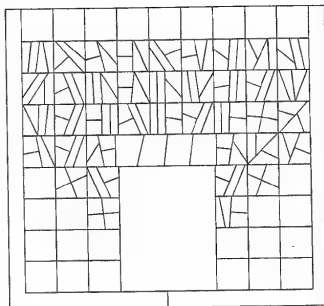


Figure 23. Fracturing in simply deformable DDA (reproduced from Lin C. *Extensions to the DDA for Jointed Rock Masses and other Blocky Systems*. PhD thesis, University of Colorado, Boulder, 1995).

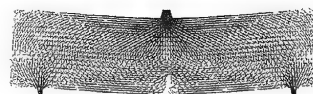


Figure 24. Fracturing of notched concrete beam modeled by jointed particulate assembly, with normal contact bond of limited strength (Irasca FPC 2D). A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

The DDA block fracturing algorithm through block centroid (Figure 23), in the context of rock fracture, comprises the Mohr–Coulomb fracturing criterion with a tension cut-off based on the stress state at the centroid, where the newly formed discontinuities are introduced and are further treated in the same way as the original discontinuity planes.

Fragmentation frameworks are also used with DEM implementations that consider particles bonded into clusters (to form more complex particle shapes), which can also be bonded into particle assemblies to represent a solid material. The bond stiffness terms are predominantly derived on the basis of equivalent continuum strain energy (Morikawa and Sawamoto, 1993; Griffiths and Mustoe, 2001). In such lattice-like models of solids, fracturing (Figure 24) is realized through breakage of interparticle lattice bonds. Typically two types of interparticle bonds are considered – a simple normal bond (or truss) and a parallel bond (or beam), which can be seen as a microscopic representation of the Cosserat continuum (see Chapter 10, Volume 2). Bond failure is typically considered on the basis of limited strength, but some softening lattice models for quasi-brittle material have also considered a gradual reduction of strength, that is softening, before a complete breakage of the bond takes place. Despite often very simple bond failure rules, bonded particulate assemblies have been shown to reproduce macroscopic manifestations of softening, dilation, and progressive fracturing.

The solution algorithm for such bonded particle assemblies remains usually an explicit time stepping scheme, that is, the overall stiffness matrix is never assembled. Steady state solutions are obtained through the dynamic relaxation (local or global damping, viscous or kinetic). A jointed particulate medium is very similar to the more recent developments in lattice models (Schlangen and van Mier, 1992; Jirasek and Bazant, 1995; D'Adetta, Ramm and Kun, 2001) for heterogeneous fracturing media. Adaptive continuum/discontinuum strategies are also envisaged, where the discrete elements are adaptively introduced into a continuum model, if and when conditions for fracturing are met (see Chapter 10, Volume 2).

7 TIME INTEGRATION — TEMPORAL DISCRETIZATION, ENERGY BALANCE, AND DISCRETE ELEMENT IMPLEMENTATION

Problems with spurious energy generation arise when applying the traditional explicit time stepping scheme to dynamic contact problems using the penalty approach. Every time the material point penetrates the boundary, it may happen that for some time that is shorter than the computational time increment, the scheme considers no contact force to resist penetration, although contact may exist for a part of the time increment. On the other hand, when the contact is released, the scheme may consider the contact force to be pushing the material point from the boundary, although there may be no penetration for certain part of the computational time increment. Consequently, some artificial spurious energy is created at every contact location and hence some form of controlling the energy balance is needed.

Local contact energy dissipation is sometimes introduced to avoid any artificial increase of energy (Munjiza, Owen and Crook, 1998). Such modified temporal operators do not affect the critical time step, and the choice of the actual computational time step can be related to a degree of numerical energy dissipation added to the system in a controlled way. For the purpose of monitoring the possible creation of spurious energy, as well as monitoring energy dissipation during fracturing, a continuous energy balance check is desired. The incorporation of softening imposes severe limits on the admissible time step.

The traditional DEM framework implies a conditionally stable explicit time stepping scheme. In the DDA context, for static problems, the resulting system of equations can be solved by any equation solver, which may be singular when blocks are separated. In order to ensure system stiffness matrix regularity in such situations, very soft spring stiffness terms are added to the block centroid deformation variables. Early realizations of the DDA framework utilized a particular type of the generalized collocation time integration scheme (Hughes, 1983) (see Chapter 5, Volume 2)

$$\begin{aligned} M\ddot{x}_{n+\theta} + C\dot{x}_{n+\theta} + Kx_{n+\theta} &= f_{n+\theta} \\ \ddot{x}_{n+\theta} &= (1-\theta)\ddot{x}_n + \theta\ddot{x}_{n+1} \\ \dot{x}_{n+\theta} &= (1-\theta)\dot{x}_n + \theta\dot{x}_{n+1} \\ x_{n+\theta} &= x_n + \theta\Delta t\dot{x}_n + (\theta\Delta t^2)[(0.5-\beta)\ddot{x}_n + \beta\ddot{x}_{n+\theta}] \\ \dot{x}_{n+\theta} &= \dot{x}_n + \theta\Delta t[(1-\delta)\ddot{x}_n + \delta\ddot{x}_{n+\theta}] \end{aligned} \quad (12)$$

In an incremental form, the recursive algorithm leads to

$$\begin{aligned} \frac{\theta M + \theta^2 \Delta t^2 \beta K}{\Delta t^2 \beta} \Delta x &= (1 - \theta) f_n + \theta f_{n+1} - K x_n \\ &+ \left[\frac{\theta M}{\Delta t \beta} + \theta^2 \Delta t K - \theta \Delta t K \right] \dot{x}_n \\ &+ \left[\frac{\theta M}{2\beta} + \frac{\theta^3 \Delta t^2 K}{2} - M - \frac{\theta^2 \Delta t^2 K}{2} \right] \ddot{x}_n \end{aligned} \quad (13)$$

In the DDA, a specific choice of the time integration parameters $\theta = 1$, $\beta = 0.5$, $\delta = 1$ is usually adopted, which represents an implicit, unconditionally stable scheme. For this choice, the effective matrix next to the acceleration vector \ddot{x}_n vanishes, hence

$$\begin{aligned} \left[\frac{2M}{\Delta t^2} + K \right] \Delta x &= \Delta f_{n+1} + \left[\frac{2M}{\Delta t} \right] \dot{x}_n \\ \hat{K} \Delta x &= \Delta f_{n+1} \end{aligned} \quad (14)$$

leading to the 'effective stiffness matrix' \hat{K} and the 'effective load vector' \hat{f} , which now include the inertia terms and the velocity at the start of the increment

$$\begin{aligned} \hat{K} &= K + \frac{2}{\Delta t^2} \int_{\Omega} T_i^T \rho T_i d\Omega \\ \Delta \hat{f}_1 &= \Delta f_1 + \frac{2}{\Delta t} \left[\int_{\Omega} T_i^T \rho T_i d\Omega \right] \dot{D}_n \\ &= \Delta f_1 + \Delta f_1 + \frac{2}{\Delta t} \left[\int_{\Omega} T_i^T \rho T_i d\Omega \right] V_n \end{aligned} \quad (15)$$

The solution for the next incremental deformation vector ΔD_{n+1} is obtained from

$$\Delta D_{n+1} = \hat{K}^{-1} \Delta \hat{f}_{n+1} \quad (16)$$

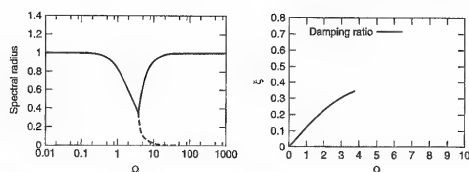


Figure 25. Spectral radius and algorithmic damping for the DDA time integration scheme (generalized Newmark algorithm $\beta = 0.5$, $\delta = 1.0$, reproduced from Doolin DM and Sitar N. Time integration in discontinuous deformation analysis. *J. Eng. Mech.* ASCE 2004; 130:249–258.)

The next time rate of deformation vector (velocity) then equals

$$V_{n+1} = \frac{2}{\Delta t} [\Delta D_{n+1}] - V_n \quad (17)$$

which is then used for the start of the next time increment.

Such a time stepping scheme can then be used to provide a dynamic response of an assembly of deformable blocks. As the effective stiffness matrix is regular due to the presence of inertia terms, block separation can be accounted for without a need to add artificial springs to block centroid variables. The time stepping procedure can also be used to obtain a steady state solution through a dynamic relaxation process. The steady state solution can be obtained by the use of the so-called *kinetic damping*, that is, by simply setting all block velocities to zero at the beginning of every increment.

It is obvious that every time integration algorithm adopted in the discrete element context (either for the combined finite/discrete element method or discontinuous deformation analysis) introduces its own characteristic numerical damping and dispersion through an apparent period elongation and amplitude decay (Chang and Acheampong, 1993). An explicit central difference scheme (used almost exclusively with the FEM/DEM method) can be viewed as a special case of the generalized Newmark algorithm with $\beta = 0$, $\delta = 0.5$ and the collocation scheme (typically used in DDA) can also be seen as a special case of the generalized Newmark algorithm, but with $\beta = 0.5$, $\delta = 1$ (Figure 25). From the analysis of spectral radius for the two recursive time integrators, it is clear that the above collocation scheme is associated with a very substantial numerical damping, which is otherwise absent in the central difference scheme. In addition, predictor-corrector or even predictor-multiple corrector schemes are adopted with granular materials (Anandarajah, 1999), in order to capture

the high-frequency events, such as the collapse of arching or avalanching.

The discrete element method simulations are computer resource intensive and therefore the need for the development of *parallel processing* procedures is clear. The explicit time integration procedure is a naturally concurrent process and can be implemented on parallel hardware in a highly efficient form (Ghaboussi, Basole and Ranjithan, 1993; Owen and Feng, 2001). The nonlinearities that are encountered in this application can also be readily incorporated. Much of the computational effort is devoted to contact detection and the search algorithms formulated principally for sequential implementation have to be restructured for parallel computation. This can cause problems for distributed memory MIMD machines, as element-to-element contact affects considerably the data communication between processors. Efficient solution procedures are employed to minimize communication requirements and maintain a load-balanced processor configuration.

Parallel implementations of DEM are typically made on both workstation clusters and multiprocessor workstations. In this way, the effort of porting software from sequential to parallel hardware is usually minimized. The program development often utilizes the sequential programming language with the organization of data structure suited to multiprocessor workstation environment, where every processor independently performs the same basic DEM algorithm on a subdomain and only communicates with other subdomains via interface data.

In view of the large mass of time-dependent data produced by a typical FEM/DEM simulation, it is essential that some means be available of continually visualizing the solution process. Visualization is particularly important in displaying the transition from a continuous to a discontinuous state. Visual representation is also used to monitor energy balance, as many discretization concepts both in space (stiffness, mass, fracturing, fragmentation, contact) and time can contribute to spurious energy imbalances (see Chapter 5, Volume 2).

8 ASSOCIATED FRAMEWORKS AND DEVELOPMENTS

This section briefly addresses some of the methods that are used to conduct an analysis of solids and structures with existing or emerging discontinuities, without necessarily accounting for the full separation of structural parts. Although such methods were initially developed as special cases of nonlinear continuum models, it may be argued that they belong to the wider category of discrete element methods, as they fundamentally address the same physical problem.

In a continuum setting, the preexisting macro discontinuities are typically accounted for through the use of interface (or joint) elements (Rots, 1988), which may be used to model crack formation as well as shearing at joints or predetermined planes of weakness. Joint planes are assumed to be of a discrete nature, their location and orientation is either given, or it becomes fixed once they are formed. The term *discrete cracking* was typically adopted, as opposed to the term *smearing cracking*, where the localized failure at an integration point level is considered in a locally averaged sense. The crack initiation across the interface element is typically driven by the tensile strength of the material (tension cutoff), and the gradual reduction of strength is modeled by some form of softening law, indicating a relationship between the tension stress across the crack and the crack opening, which may include a softening law usually controlled by the fracture energy release rate (see Chapter 10 and Chapter 11 of Volume 2). Similarly, in modeling shear response, the Coulomb criterion is usually adopted, with the gradual decohesion at the interface. Natural extension of the above simple concepts leads to the interface material models that account for a combination of cracking and Coulomb friction, which have been formulated as two-parameter failure surfaces in the context of computational plasticity.

The rigid bodies spring model, RBSP (Kawai, 1977), was earlier proposed as a generalized limit plastic analysis framework. The discrete concept of a discontinuity is present as well, but the deformability of the material between the discontinuities is ignored. Structures are modeled as assemblies of rigid blocks connected to one another by normal and tangential springs on their interfaces. The purpose of the analysis is to start from a state where the structure is represented by a connected assembly of rigid blocks and where the progressive failure is modeled by developing discontinuities, through emerging cracks and slipping at the rigid blocks interfaces.

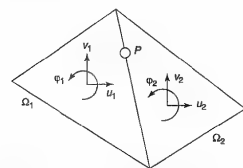


Figure 26. Two rigid blocks with an elastic interface contact in RBSP.

In the RBMS, the stiffness matrix is obtained by considering rigid bodies to be connected by distributed normal and tangential springs of the values k_n and k_t per unit length. The rigid displacement field within an arbitrary block (Figure 26) is expressed in terms of the centroid displacements and rotation $(u, v, \theta)^T$.

The constitutive relation between the traction components and the relative displacement at the location P in plane stress can be written as

$$\sigma = D\delta \quad \sigma = [\sigma_n, \tau_t]^T \quad (18)$$

$$\text{with } D = \begin{bmatrix} k_n & 0 \\ 0 & k_t \end{bmatrix} \quad k_n = \frac{(1-\nu)E}{h(1-2\nu)(1+\nu)}$$

$$k_t = \frac{E}{h(1+\nu)} \quad (19)$$

where h is the sum of shortest distances h_1 and h_2 between the two block centroids to the contact line.

Generalization to 3-D situation is conceptually straightforward, and the method can clearly be interpreted as a similar method to the finite element method with joint (interface) elements of zero thickness, the only difference coming from the assumption that the overall elastic behavior is represented only by the distributed stiffness springs along interfaces. The RBMS has been successfully used in the nonlinear (failure) analysis of the plain concrete and reinforced concrete structures.

There are several other modeling frameworks that fall into the category of discrete element methods, for example, modified distinct element method, MDEM (Meguro and Tagel-Din, 1999). The non smooth contact dynamics method, NSCD (Moreau, 1999; Jean, 1999; Jean, Acary and Monerie, 2001), is related to both the combined FEM/DEM and the DDA, but it comprises significant differences, as the unilateral Signorini condition and the dry friction Mohr-Coulomb are adopted without resorting to smooth regularization. In the context of multiple bodies contact (i.e. fragmentation or granular flow), with discontinuities in the field of velocities, NSCD follows the argument that it is not possible to define the acceleration as a usual derivative of a smooth function. Instead of using the regularized form of dynamic equations of equilibrium, the nonsmooth time discretized form of dynamic equations is obtained by integrating the dynamic equation, so that the velocities become the primary unknowns

$$M(\dot{x}_{i+1} - \dot{x}_i) = \int_{\Gamma_i}^{t_{i+1}} F(t, x, \dot{x}) dt + \int_{t_i}^{t_{i+1}} r dt \quad (20)$$

$$x_{i+1} = x_i + \int_{t_i}^{t_{i+1}} \dot{x} dt$$

Another feature of the NSCD is that the force impulse on the RHS is split into two parts, where the first part $\int_{t_i}^{t_{i+1}} F(t, x, \dot{x}) dt$ (which excludes the contact forces) is considered continuous, whereas the second part $\int_{t_i}^{t_{i+1}} r dt$ (representing contact forces contribution to force impulse) is replaced by the mean value impulse $r_{i+1} = (1/\Delta t) \int_{t_i}^{t_{i+1}} r dt$ over the finite time interval. In physical terms, this implies that the full contact interaction history is only accounted for in an averaged sense over the time interval, discarding the details, which are either deemed impossible to characterize owing to insufficient data available, or are deemed inconsequential in terms of having a significant effect on the overall solution.

Different time stepping algorithms are possible depending on the approximation adopted and on the nature of the constitutive model for the block material. A low-order implicit time stepping scheme, typically the backward Euler scheme is used. Resolution for contact kinematics then considers the relationship between the relative velocities of contacting bodies and the mean value impulse at discrete contact locations

$$\dot{x}_{rel} = \dot{x}_{rel-sec} + K_{eff}^{-1} \Delta T r_{i+1} \quad (21)$$

where the first part represents the 'free' relative velocity (without the influence of contact forces) and the second part comprises 'corrective' velocities emanating from contacts. The actual algorithm is realized in a similar manner to the closest point projection stress return schemes in computational plasticity – a 'free predictor' for relative velocities is followed by iterations to obtain 'iterative corrector' values for mean value impulses, such that the inequality constraints (Signorini and Mohr-Coulomb) are satisfied in a similar manner as the plastic consistency condition is iteratively satisfied in computational plasticity algorithms. The admissible domains in contact problems are generally nonconvex and it is argued that it is necessary to treat the contact forces in a fully implicit manner, whereas other forces can be considered either explicitly or implicitly (Kane *et al.*, 1999), leading to either implicit/implicit or explicit/implicit algorithms.

Furthermore, there is a degree of commonality of novel ideas in terms of describing the block deformability in discrete element methods and novel developments in the continuum-based techniques. The *manifold method* (Shi, 1997; Chen, Ohnishi and Ito, 1997) advocates similar ideas as the ones advocated in the *meshless* (Belytschko *et al.*, 1996) or the *partition of unity* (Melenk and Babuška, 1996) methods (see Chapter 10, Volume 2). Similar to the meshless methods (see Chapter 10, this Volume), the manifold method identifies the cover displacement function C_i and the cover weighting function w_i , where the geometry of

the actual blocks Ω_i is utilized for numerical integration purposes over the background grid. The treatment of any emerging discontinuities is envisaged by introducing the concept of effective cover regions, where there is a need to introduce n independent covers, if a cover intersects n disconnected domains. These concepts point to a range of possibilities in the simulation of progressive discontinuities in quasi-brittle materials.

Discontinuous modeling frameworks are also increasingly moving toward formulations and applications in multifield and multiphysics problems, in particular, in the area of the coupled fluid flow in discontinuous, jointed media. It is believed that discontinuous modeling frameworks have a bright and exciting future, especially in the context of fragmentation and in the microscopic simulation of the behavior of heterogeneous materials, where the notion that simple constitutive laws at the micro, meso, or nano level generate manifestations of complex macroscopic behavior, such as plasticity or fracture (Cundall, 2002), is fundamentally different to the top-down approach of the nonlinear FEM. Increased computing power and efficient contact detection algorithms will not only allow modeling of progressive fracturing, including fragmented state, but will also allow for the further development and enhancement of discrete microstructural models of material behavior where the continuum concept may be abandoned and an internal length scale may be intrinsically incorporated into the model. Moreover, the large scale simulations with adaptive multiscale material models, where different regions or domains are accounted for at a different scale of observation are governed by noncontinuum laws of physics, seem possible in a not-too-distant future.

REFERENCES

- Anandarajah A. Multiple time-stepping for the discrete element analysis of colloidal particles. *Powder Technol.* 1999; **106**: 132–141.
- Barbosa R and Ghaboussi J. Discrete Finite Element Method. *Eng. Comput.* 1992; **9**:253–266.
- Belytschko T, Krongauz Y, Organ D, Fleming M and Krysl P. Meshless methods: an overview and recent developments. *Comput. Methods Appl. Mech. Eng.* 1996; **139**:3–47.
- Bonet J and Peraire J. An alternating digital tree (ADT) for 3D geometric searching and intersection problems. *Int. Journal Num. Meth. Engng.* 1991; **31**:1–17.
- Chang CS and Acheampong KB. Accuracy and stability for static analysis using dynamic formulation in discrete element methods. In *2nd International Conference on Discrete Element Methods*, Williams J and Mustoe GGW (eds). MIT IESL Publication, 1993.
- Chen G, Ohnishi Y and Ito T. Development of high order manifold method. In *ICADD-2 Conference on Analysis of Discontinuous Deformation*, Ohnishi Y (ed.). Kyoto University, 1997.
- Cleary PW. Large scale industrial DEM modelling. In *Discrete Element Methods: Numerical Modeling of Discontinua, 3rd International Conference on Discrete Element Methods*, Cook BK and Jensen PJ (eds). ASCE Geotechnical Special Publication No. 117, The Geo-Institute of the American Society of Civil Engineers, 2002.
- Cundall PA. A computer model for simulating progressive large scale movements in blocky rock systems. *Proceedings Symposium Int Soc Rock Mech*, Nancy France, Vol 1, Paper II-8, 1971.
- Cundall PA and Strack ODL. A discrete numerical model for granular assemblies. *Geotechnique* 1979; **29**:47–65.
- Cundall PA. Formulation of a three-dimensional distinct element model – Part I: A scheme to detect and represent contacts in a system composed of many polyhedral blocks. *Int. J. Rock Mech., Min. Sci. Geomech. Abstr.* 1988; **25**:107–116.
- Cundall PA. Numerical experiments on localization in frictional materials. *Ingenieur-Archiv* 1989; **59**:148–159.
- Cundall PA. A discontinuous future for numerical modelling in soil and rock. In *Discrete Element Methods: Numerical Modeling of Discontinua, 3rd International Conference on Discrete Element Methods*, Cook BK and Jensen PJ (eds). ASCE Geotechnical Special Publication No. 117, The Geo-Institute of the American Society of Civil Engineers, 2002.
- D'Alesta GA, Ramen E and Kun F. Fracture simulations of cohesive frictional materials by discrete element models. In *ICADD-4, 4th Conference on Analysis of Discontinuous Deformation*, Bićanić N (ed.). University of Glasgow: Glasgow, 2001; 1–494.
- De Borst R. Some recent issues in computational failure mechanics. *Int. J. Numer. Methods Eng.* 2001; **52**:63–96.
- Dociu DM and Sitar N. Time integration in discontinuous deformation analysis. *J. Eng. Mech. ASCE* 2004; **130**:249–258.
- Feng YT and Owen DRJ. An augmented spatial digital tree algorithm for contact detection in computational mechanics. *Int. J. Numer. Methods Eng.* 2002a; **55**:159–176.
- Feng YT and Owen DRJ. An energy based corner to corner contact algorithm. In *Discrete Element Methods: Numerical Modeling of Discontinua, 3rd International Conference on Discrete Element Methods*, Cook BK and Jensen PJ (eds). ASCE Geotechnical Special Publication No. 117, The Geo-Institute of the American Society of Civil Engineers, 2002b.
- Ghaboussi J. Fully deformable discrete element analysis using a finite element approach. *Int. J. Comput. Geotech.* 1988; **5**:175–195.
- Ghaboussi J, Basole MM and Ranjithan S. Three dimensional discrete element analysis on massively parallel computers. In *2nd International Conference on Discrete Element Methods*, Williams J and Mustoe GGW (eds). MIT IESL Publication, 1993.
- Goodman RE, Taylor RL and Brekke T. A model for mechanics of jointed rock. *J. Soil Mech. Found. Div. Proc. ASCE* 1968; **94**:SM3.
- Griffiths DV and Mustoe GGW. Modelling of elastic continua using a Grillage of structural elements based on discrete

- element concepts. *Int. J. Numer. Methods Eng.* 2001; **50**: 1759–1775.
- Groger T, Tuzun U and Heyes D. Shearing of wet particle systems discrete element simulations. In *1st International PFC Symposium*, Itasca, 2002.
- Hocking G. The DEM for analysis of fragmentation of discontinua. In *1st U.S. Conference on Discrete Element Methods (DEM)*, Mustoe GGW, Henriksen M and Hutelmaier HP (eds). Colorado School of Mines Press, Golden, 1989.
- Hocking G, Williams JR and Mustoe GGW. Dynamics analysis for three dimensional contact and fracturing of multiple bodies. In *NUMETA 87, Numerical Methods in Engineering, Theory and Applications*, Pande GN and Middleton J (eds). Martinus Nijhoff Publishers, 1987.
- Hogue C. Shape representation and contact detection for discrete element simulations of arbitrary geometries. *Eng. Comput.* 1998; **3**:374–390.
- Hughes TJR. Analysis of transient algorithms with particular reference to stability behaviour. In *Computational Methods for Transient Analysis*, Belytschko T and Hughes TJR (eds). Elsevier Science Publishers, 1983.
- Jean M. The non-smooth contact dynamics method. *Comput. Methods Appl. Mech. Eng.* 1999; **177**:235–257.
- Jean M, Acary V and Monerie Y. Non-smooth contact dynamics approach of cohesive materials. *Philos. Trans. R. Soc. London* 2001; **359**:1–22.
- Jirasek M and Bazant ZP. Macroscopic fracture characteristics of random particle systems. *Int. J. Fracture* 1995; **69**:201–228.
- Kane C, Repetto EA, Ortiz M and Marsden JE. Finite element analysis of non smooth contact. *Comput. Methods Appl. Mech. Eng.* 1999; **180**:1–26.
- Kawai T. New element models in discrete structural analysis. *J. Soc. Naval Arch. Jpn.* 1977; **141**:187–193.
- Ke TC. The issue of rigid body rotation in DDA. In *First Intl Forum on DDA and Simulation of Discontinuous Media*, Salami MR and Banks D (eds). TSI Press, 1996.
- Kremmer M and Favier JF. A method for representing boundaries in discrete element modelling – part I Geometry and contact detection. *Int. J. Numer. Methods Eng.* 2001a; **51**:1407–1421.
- Kremmer M and Favier JF. A method for representing boundaries in discrete element modelling – part II Kinematics. *Int. J. Numer. Methods Eng.* 2001b; **51**:1423–1436.
- Kuhn M. A flexible boundary for three dimensional DEM particle assemblies. In *2nd International Conference on Discrete Element Methods*, Williams J and Mustoe GGW (eds). MIT IESL Publication, 1993.
- Lazarević D and Dvornik J. Selective time steps in predictor-corrector methods applied to discrete dynamic models of granular materials. In *ICADD-4, 4th Conference on Analysis of Discontinuous Deformation*, Bićanić N (ed.). University of Glasgow, 2001.
- Lemos JV. Assessment of the ultimate load of masonry arch using discrete elements. In *Comp Meth in Struct Masonry 3*, Middleton J and Pande G (eds). Books and Journals International, 1995.
- Lin C. *Extensions to the DDA for Jointed Rock Masses and other Blocky Systems*. PhD thesis, University of Colorado, Boulder, 1995.
- Lin X and Ng TT. Contact detection algorithm for three dimensional ellipsoids in discrete element modelling. *Int. J. Numer. Anal. Methods Geomech.* 1995; **19**:653–659.
- Ma MY, Zaman M and Zhou JH. Discontinuous deformation analysis using the third order displacement function. In *First Intl Forum on DDA and Simulation of Discontinuous Media*, Salami MR and Banks D (eds). TSI Press, 1996.
- Melenk JM and Babuška I. The partition of unity finite element method – basic theory and applications. *Comput. Methods Appl. Mech. Eng.* 1996; **139**:289–314.
- McLaughlin MM and Sitar N. Rigid body rotations in DDA. In *First Intl Forum on DDA and Simulation of Discontinuous Media*, Salami MR and Banks D (eds). TSI Press, 1996.
- McNamara S and Young WR. Inelastic collapse in two dimensions. *Phys. Rev. E* 1994; **50**:28–31.
- Moreau JJ. Numerical aspects of the sweeping process. *Comput. Methods Appl. Mech. Eng.* 1999; **177**:329–349.
- Morioka H and Sawamoto Y. Local fracture analysis of a reinforced slab by the discrete element method. In *2nd International Conference on Discrete Element Methods*, Williams J and Mustoe GGW (eds). MIT IESL Publication, 1993.
- Müller D. *Techniques Informatiques Efficaces pour la Simulation de vMieux Granulaires par des Methodes d'Elements Distincts*, These 1545, EFF Lausanne, 1996.
- Munjiza A and Andrews KRF. NBS contact detection algorithm for bodies of similar size. *Int. J. Numer. Methods Eng.* 1998; **43**:131–149.
- Munjiza A, Owen DRJ and Bićanić N. A combined finite/discrete element method in transient dynamics of fracturing solids. *Eng. Comput.* 1995; **12**:145–174.
- Munjiza A, Owen DRJ and Crook AJL. An $M(M^{-1}K)^m$ proportional damping in explicit integration of dynamic structural systems. *Int. J. Numer. Methods Eng.* 1998; **41**:1277–1296.
- Mustoe GGW, Williams JR and Hocking G. Penetration and fracturing of brittle plates under dynamic impact. In *NUMETA 87, Numerical Methods in Engineering, Theory and Applications*, Pande GN and Middleton J (eds). Martinus Nijhoff Publishers, 1987.
- Ng TT. Hydrostatic boundaries in discrete element methods. In *Discrete Element Methods: Numerical Modeling of Discontinua, 3rd International Conference on Discrete Element Methods*, Cook BK and Jensen PJ (eds). ASCE Geotechnical Special Publication No. 117, The Geo-Institute of the American Society of Civil Engineers, 2002.
- O'Connor R, Gill JJ and Williams JR. A linear complexity contact detection algorithm for multi-body simulation. In *2nd International Conference on Discrete Element Methods*, Williams J and Mustoe GGW (eds). MIT IESL Publication, 1993.
- Owen DRJ, Perić D, de Souza Neto EA, Crook AJL, Yu J and Klerck PA. Computational strategies for discrete systems and multi-fracturing materials. In *ECCEM European Conference on Computational Mechanics*, München, 1999.
- Owen DRJ and Feng YT. Parallelised finite/discrete element simulation of multi fracture solids and discrete systems. *Eng. Comput.* 2001; **18**:557–576.
- Pentland AP and Williams JR. Good Vibrations: Modal Dynamics for Graphics and Animation. *ACM Computer Graphics*, 1989; **23**(3): 215–222.
- Perkins E and Williams JR. A fast contact detection algorithm insensitive to object sizes. *Eng. Comput.* 2001; **18**(1–2):48–61.
- Petrinić N. *Aspects of Discrete Element Modelling Involving Facet-to-Facet Contact Detection and Interaction*, PhD thesis, University of Wales, Swansea, 1996.
- Rots J. Computational Modelling of Concrete Fracture. *Heron*, 1988; **34**:81–88.
- Schlangen EJ and van Mier J. Experimental and numerical analysis of micromechanisms of fracture of cement based composites. *Cement Concrete Compos.* 1992; **6**:1–59.
- Shi GH. *Discontinuous Deformation Analysis – a New Numerical Model for Statics and Dynamics of Block Systems*. PhD thesis, University of California, Berkeley, 1988.
- Shi GH. Numerical manifold method. In *ICADD-2 Conference on Analysis of Discontinuous Deformation*, Ohmishi Y (ed.). Kyoto University, 1997.
- Shi GH and Goodman RE. A new concept for support of underground and surface excavations in discontinuous rocks based on a keystone principle. *22nd U.S. Symposium on Rock Mechanics*, MIT, 1981.
- Taylor LM and Preece DS. Simulation of blasting induced rock motion using spherical element models. In *1st U.S. Conference on Discrete Element Methods (DEM)*, Mustoe GGW, Henriksen M and Hutelmaier HP (eds). Colorado School of Mines Press, Golden, 1989.
- Thomton C. Applications of DEM to process engineering problems. *Eng. Comput.* 1992; **9**:289–197.
- Ting JM. A robust algorithm for ellipse based discrete element modelling of granular materials. *Comput. Geotech.* 1992; **13**:175–186.
- Vu-Quoc L, Zhang X and Walton OR. A 3-discrete-element method for dry granular flows of ellipsoidal particles. *Comput. Methods Appl. Mech. Eng.* 2000; **187**:483–528.
- Wait R. Discrete element models of particle flows. *Math. Modell. Anal.* 2001; **6**:156–164.
- Williams JR and Mustoe GGW. Modal methods for the analysis of discrete systems. *Comput. Geotech.* 1987; **4**:1–19.
- Williams JR and O'Connor R. Discrete element simulation and the contact problem. *Archiv. Comput. Methods Eng.* 1999; **6**:279–304.
- Williams JR and Pentland AP. Superquadrics and modal dynamics for discrete elements in interactive design. *Eng. Comput.* 1992; **9**:115–127.
- Williams JR, Hocking G and Mustoe GGW. The theoretical basis of the discrete element method. *NUMETA 85, Numerical Methods of Engineering, Theory and Applications*, A.A. Balkema, Rotterdam, 1985.
- Williams JR, Mustoe GGW and Hocking G. Three dimensional analysis of multiple bodies including automatic fracturing. In *COMPLAS, 1st International Conference on Computational Plasticity*, Owen DRJ, Hinton E and Onate E (eds). Pineridge Press, 1987.
- Clearly PW. DEM simulation of industrial particle flows: case studies of dragline excavators, mixing in tumblers and centrifugal mills. *Powder Technol.* 2000; **109**:83–104.
- Cook BK and Jensen PJ (eds). Discrete element methods: numerical modeling of discontinua. *3rd International Conference on Discrete Element Methods*, ASCE Geotechnical Special Publication No. 117, The Geo-Institute of the American Society of Civil Engineers, 2002.
- Cundall PA. UDEC – A Generalised Distinct Element Program for Modelling Jointed Rock. US Army European Research Office, NTIS AD-A087-610/2, 1987.
- Cundall PA and Hart RD. Numerical modeling of discontinua. In *1st U.S. Conference on Discrete Element Methods (DEM)*, Mustoe GGW, Henriksen M and Hutelmaier HP (eds). Colorado School of Mines Press, Golden, 1989.
- Cundall PA and Hart RD. Numerical modelling of discontinua. *Eng. Comput.* 1992; **9**:101–114.
- Fortin J and Hjjaj Mand de Saxe G. An improved discrete element method based on a variational formulation of the frictional contact law. *Comput. Geotech.* 2002; **29**: 609–640.
- Han K, Perić D, Crook AJL and Owen DRJ. A combined finite/discrete element simulation of shot peening processes – Parts I and II – 2D and 3D interaction laws. *Eng. Comput.* 2000; **17**:593–619, 680–702.
- Holst JM, Rotter JM, Ooi JY and Rong GH. Numerical modelling of Silo filling – discrete element analysis. *ASCE J. Eng. Mech.* 1999; **125**:94–110.
- Hogue C and Newland D. Efficient computer simulation of moving granular particles. *Powder Technol.* 1994; **78**:51–66.
- Meguro K and Tagel-Din H. A new simplified and efficient technique for fracture behaviour analysis of concrete structures. In *FRAMCOS-3 International Conference Fracture Mechanics of Concrete Structures*, Academic Publishers: Freiburg, 1999.
- Munjiza A and Andrews KRF. Penalty function method for combined finite-discrete element systems comprising large number of separate bodies. *Int. J. Numer. Methods Eng.* 2000; **49**:1377–1396.
- Munjiza A and John NWM. Mesh size sensitivity of the combined FEM/DEM fracture and fragmentation algorithms. *Eng. Fracture Mech.* 2002; **69**:281–295.
- Munjiza A, Latham JP and John NWM. 3D dynamics of discrete element systems comprising irregular discrete elements – integration solution for finite rotations in 3D. *Int. J. Numer. Methods Eng.* 2003; **56**:35–55.
- Mustoe GGW, Henriksen M and Hutelmaier HP (eds). *Proceedings of the 1st U.S. Conference on Discrete Element Methods (DEM)*, Colorado School of Mines, Golden, CO, 1989.
- Williams JR and Mustoe GGW (eds). *2nd International Conference on Discrete Element Methods (DEM)*, Massachusetts Institute of Technology, MIT IESL Publications, 1993.

Chapter 12

Boundary Element Methods: Foundation and Error Analysis

G. C. Hsiao¹ and W. L. Wendland²

¹University of Delaware, Newark, DE, USA

²Universität Stuttgart, Stuttgart, Germany

| | |
|-------------------------------|-----|
| 1 Introduction | 339 |
| 2 Boundary Integral Equations | 340 |
| 3 Variational Formulations | 347 |
| 4 The Galerkin-BEM | 358 |
| 5 The Role of Sobolev Index | 366 |
| 6 Concluding Remarks | 371 |
| Acknowledgments | 371 |
| References | 371 |
| Further Reading | 373 |

1 INTRODUCTION

In essence, the boundary element method (BEM) may be considered as an application of finite element method (FEM), designed originally for the numerical solutions of partial differential equations (PDE) in the domains, to the boundary integral equations (BIE) on closed boundary manifolds. The terminology of BEM originated from the practice of discretizing the boundary manifold of the solution domain for the BIE into boundary elements, resembling the term of finite elements in FEM. As in FEM, in the literature, the use of the terminology *boundary element* is in two different contexts: the boundary manifolds are decomposed into boundary elements, which

are geometric objects, while the boundary elements for approximating solutions of BIEs are actually the finite element functions defined on the boundaries. Looking through the literature, it is difficult to trace back one fundamental research paper and the individuals who were responsible for the historical development of the BEM. However, from the computational point of view, the work by Hess and Smith deserves mention as one of the cornerstones of BEM. In their 1966 paper (Hess and Smith, 1966), boundary elements (or rather surface elements) have been used to approximate various types of bodies and to calculate the potential flow about arbitrary bodies. On the other hand, the paper by Nedelec and Planchard (1973) may be considered as a genuine boundary element paper with respect to the variational formulation of BIEs. Other early contributions to the boundary element development in the 1960s and 1970s from the mathematical point of view include Fichera (1961), Wendland (1965, 1968), MacCamy (1966), Mikhlin (1970), Hsiao and MacCamy (1973), Stephan and Wendland (1976), Jaswon and Symm (1977), LeRoux (1977), Nedelec (1977), and Hsiao and Wendland (1977), to name a few.

The BEM has received much attention and gained wide acceptance in recent years. From 1989 to 1995, the German Research Foundation DFG installed a Priority Research Program 'Boundary Element Methods', and the final report appeared as a book (see Wendland, 1997). There has been an increasing effort in the development of efficient finite element solutions of BIEs arising from elliptic boundary value problems (BVP). In fact, nowadays, the term BEM denotes any 'efficient method' for the approximate numerical solution of these boundary integral equations.

One of the distinct features of the method is that the approximate solution of the boundary value problem via BEM will always satisfy the corresponding PDEs exactly in the domain and is characterized by a finite set of parameters on the boundary.

As the classical integral equation method for numerical solutions to elliptic BVPs, central to the BEM is the reduction of BVPs to the equivalent integral equations on the boundary. This boundary reduction has the advantage of diminishing the number of space dimension by 1 and the capability to handle problems involving infinite domains. The former leads to an appreciable reduction in the number of algebraic equations generated for the solutions, as well as a much simplified data representation. On the other hand, it is well known that elliptic BVPs may have equivalent formulations in various forms of BIEs. This provides a great variety of versions for BEMs. However, irrespective of the variants of the BEMs and the particular numerical implementation chosen, there is a common mathematical framework into which all these BEMs may be incorporated. This chapter addresses the fundamental issues of this

common mathematical framework and is devoted to the mathematical foundation underlying the BEM techniques.

Specifically, this chapter will give an expository introduction to the Galerkin-BEM for elliptic BVPs from the mathematical point of view. Emphasis will be placed on the variational formulations of the BIEs and the general error estimates for the approximate solutions in appropriate Sobolev spaces. A classification of BIEs will be given on the basis of the Sobolev index. The simple relations between the variational formulations of the BIEs and the corresponding PDEs under consideration will be indicated. Basic concepts such as stability, consistency, and convergence, as well as the condition numbers and ill-posedness, will be discussed by using elementary examples.

Figure 1 is a sketch of the general procedure for approximating the solutions of a BVP via the boundary element methods. In the remaining sections, we will discuss the topics by following the procedure up to including the asymptotic error estimates and end the chapter with further reading materials with reference to other topics that are not included here because of the limitation of the length of the chapter.

2 BOUNDARY INTEGRAL EQUATIONS

In this section, basic BVPs in elasticity will be presented as the mathematical model problems to illustrate the general procedure for the BEM given in the introduction. We begin with the reduction of boundary value problems to various boundary integral equations. The concepts concerning fundamental solutions, Green's representation formula, Cauchy data, and four basic boundary integral operators will be introduced. Various numerical schemes for the derived BIEs will be discussed, and the corresponding algebraic equations will be formally obtained in terms of boundary elements.

These model problems in elasticity are particularly educational in the sense that solutions of the equivalent BIEs are not always unique, even though the original BVP is uniquely solvable. As will be seen, the rigid motions will play an important role in circumventing this difficulty. Throughout the chapter, $\Omega \subset \mathbb{R}^n$, $n = 2$ or 3 , denotes a bounded domain with smooth boundary Γ and $\Omega^c = \mathbb{R}^n \setminus \Omega$, the exterior domain.

2.1 Boundary value problems

In linear elasticity for isotropic materials, the governing equations are

$$-\Delta^* \mathbf{U} = \mathbf{f} \quad \text{in } \Omega(\text{or } \Omega^c) \quad (1)$$

where

$$\Delta^* \mathbf{U} := \mu \Delta \mathbf{U} + (\mu + \lambda) \operatorname{grad} \operatorname{div} \mathbf{U}$$

is the Lamé operator, \mathbf{U} the unknown displacement field, and \mathbf{f} a given body force. Here, μ and λ are given constants such that $\mu > 0$ and $\lambda > -(2/n)\mu$. These are so-called Lamé constants, which characterize the elastic material. We consider four fundamental BVPs, the interior and exterior displacement, and traction problems. The displacement problem is a BVP of the Dirichlet type in which the boundary displacement

$$\mathbf{U}|_{\Gamma} = \boldsymbol{\phi} \quad \text{on } \Gamma \quad (2)$$

is prescribed, while the traction problem is of the Neumann type with the traction boundary condition

$$\mathbf{T}\mathbf{U}|_{\Gamma} = \boldsymbol{\psi} \quad \text{on } \Gamma \quad (3)$$

Here, \mathbf{T} is the traction operator defined by

$$\mathbf{T}\mathbf{U}|_{\Gamma} := \left(\lambda(\operatorname{div} \mathbf{U})\mathbf{n} + 2\mu \frac{\partial \mathbf{U}}{\partial \mathbf{n}} + \mu \mathbf{n} \times \operatorname{curl} \mathbf{U} \right) \Big|_{\Gamma} \quad (4)$$

for $n = 3$, which reduces to the case for $n = 2$ by setting $U_3 = 0$ and the third component of the normal $n_3 = 0$. Here, and in what follows, \mathbf{n} is always the exterior unit normal to Ω . For the exterior BVPs, we require, as usual, appropriate growth conditions, which will be specified later.

To reduce the BVPs to BIEs, one needs the knowledge of the fundamental solution $E(x, y)$ for the PDE (1), a distribution satisfying

$$-\Delta^* E(x, y) = \delta(|x - y|)I$$

with the Dirac- δ function and identity matrix I , which can be derived by using the Fourier transformation. A simple calculation shows that

$$E(x, y) = \frac{\lambda + 3\mu}{4\pi(n-1)\mu(\lambda + 2\mu)} \times \left\{ \gamma(x, y)I + \frac{\lambda + \mu}{\lambda + 3\mu} \frac{1}{|x - y|^n} (x - y)(x - y)^T \right\} \quad (5)$$

a matrix-valued function, where

$$\gamma(x, y) = \begin{cases} -\log|x - y| & \text{for } n = 2 \\ \frac{1}{|x - y|} & \text{for } n = 3 \end{cases}$$

(In fact, these are the fundamental solutions for the Laplacian.) In the so-called *direct approach* for the BEM, the

BIEs are based on the Green representation formula, which in elasticity is termed the *Betti-Somigliana representation formula* for the solutions of the BVPs under consideration. For interior problems, we have the representation

$$\mathbf{U}(x) = \int_{\Gamma} E(x, y) \mathbf{T}\mathbf{U}(y) ds_y - \int_{\Gamma} (T_y E(x, y))^T \mathbf{U}(y) ds_y - \int_{\Omega} E(x, y) \mathbf{f}(y) dy \quad (6)$$

for $x \in \Omega$. The subscript y in $T_y E(x, y)$ denotes differentiations in (6) with respect to the variable y .

The last term in the representation (6) is the volume potential due to the given body force \mathbf{f} defining a particular solution \mathbf{u}_p of (1). For linear problems, we may decompose the solution in the form

$$\mathbf{U} = \mathbf{u}_p + \mathbf{u}$$

where \mathbf{u} now satisfies the homogeneous equation (1) with $\mathbf{f} = 0$ and has a representation from (6) in the form

$$\mathbf{u}(x) = V\boldsymbol{\sigma}(x) - W\boldsymbol{\phi}(x), \quad x \in \Omega \quad (7)$$

Here, V and W are respectively the *simple*- and *double*-layer potentials defined by

$$V\boldsymbol{\sigma}(x) := \int_{\Gamma} E(x, y) \boldsymbol{\sigma}(y) ds_y$$

$$W\boldsymbol{\phi}(x) := \int_{\Gamma} (T_y E(x, y)) \boldsymbol{\phi}(y) ds_y$$

for $x \in \Omega$; and the boundary charges $\boldsymbol{\phi}(x) = \mathbf{u}(x)|_{\Gamma}$, $\boldsymbol{\sigma}(x) = \mathbf{T}\mathbf{u}(x)|_{\Gamma}$ are the *Cauchy data* of the solution \mathbf{u} to the homogeneous equation

$$\Delta^* \mathbf{u} = 0 \quad \text{in } \Omega \quad (8)$$

Because of the above decomposition, in the following sections we shall consider, without loss of generality, only the homogeneous equation (8).

For exterior problems, the representation formula for \mathbf{u} needs to be modified by taking into account the growth conditions at infinity. The appropriate growth conditions are

$$\mathbf{u}(x) = -E(x, 0)\boldsymbol{\Sigma} + \omega(x) + O(|x|^{1-n}) \quad \text{as } |x| \rightarrow \infty \quad (9)$$

where $\omega(x)$ is a rigid motion defined by

$$\omega(x) = \begin{cases} \mathbf{a} + \mathbf{b}(-x_2, x_1)^T & \text{for } n = 2 \\ \mathbf{a} + \mathbf{b} \times (x_1, x_2, x_3)^T & \text{for } n = 3 \end{cases} \quad (10)$$

Here, \mathbf{a} , \mathbf{b} , and $\boldsymbol{\Sigma}$ are constant vectors related to the translation, rotation, and total boundary forces respectively.

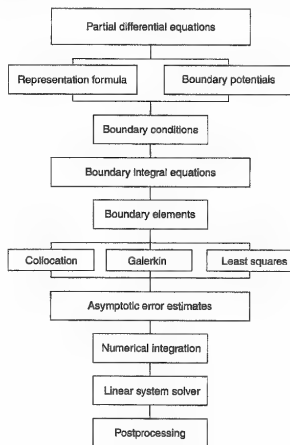


Figure 1. A schematic procedure for boundary element methods.

The total surface forces Σ are generally specified and are related to the rigid motion ω . The latter may or may not be specified, depending on the dimension and the problems under consideration. The representation formula for solutions of

$$\Delta^* u = 0 \text{ in } \Omega^c \quad (11)$$

together with the growth condition (9) now has the form

$$u(x) = -V\sigma(x) + W\varphi(x) + \omega(x) \quad (12)$$

for $x \in \Omega^c$, with $\varphi = u|_\Gamma$ and $\sigma = T u|_\Gamma$, being the Cauchy data and

$$\Sigma = \int_\Gamma \sigma \, ds \quad (13)$$

We remark that (7), (12), and (13) are the basic representations for the interior and exterior problems in the *direct approach*. Alternatively, there is a so-called *indirect approach*; one seeks the solutions of BVPs in the form of either simple- or double-layer potentials. Then the unknown boundary charges (or densities) are no longer the Cauchy data but the jumps of the corresponding Cauchy data across the boundary. Nevertheless, one may arrive at similar BIEs as in the direct approach but with a simplified given right-hand side, which, in contrast to the direct approach, may or may not always be in the range of the corresponding boundary integral operators. Moreover, desired physical quantities such as the Cauchy data need to be calculated from the density functions in some postprocessing procedure.

2.2 Basic boundary integral operators

We begin with the interior problems. Applying the trace and the traction operator T in (4) to both sides of the representation formula (7), we obtain the overdetermined system of BIEs on Γ

$$\varphi = (\tfrac{1}{2}I - K)\varphi + V\sigma \quad (14)$$

$$\sigma = D\varphi + (\tfrac{1}{2}I + K')\sigma \quad (15)$$

for the Cauchy data φ, σ of the solution u . Here, V, K, K' , and D are the four basic boundary integral operators, which are generally known respectively as the *simple-layer*, *double-layer*, *transpose of the double-layer*, and *hyperangular* boundary integral operators in potential theory. They are defined by

$$V\sigma(x) := \lim_{\Omega z \rightarrow x \in \Gamma} V\sigma(z) \quad (16)$$

$$K\varphi(x) := \lim_{\Omega z \rightarrow x \in \Gamma} W\varphi(z) + \tfrac{1}{2}\varphi(x) \quad (17)$$

$$K'\sigma(x) := \lim_{\Omega z \rightarrow x \in \Gamma} T_z V\sigma(z) - \tfrac{1}{2}\sigma(x) \quad (18)$$

$$D\varphi(x) := - \lim_{\Omega z \rightarrow x \in \Gamma} T_z W\varphi(z) \quad (19)$$

for $x \in \Gamma$, where, in the neighborhood of Γ , the traction operator T_z is defined by

$$T_z v(z) := [\lambda(\operatorname{div}_z v(z)) + 2\mu(\operatorname{grad}_z v(z))] \cdot n_x + \mu n_x \times (\operatorname{curl}_z v(z))$$

for $z \in \Omega$ or $z \in \Omega^c$. From these definitions, we have the following classical results.

Lemma 1. Let $\Gamma \in C^2$ and let $\varphi \in C^\alpha(\Gamma)$, $\sigma \in C^\alpha(\Gamma)$ with $0 < \alpha < 1$. Then the limits in (16) to (18) exist uniformly with respect to all $x \in \Gamma$ and all φ and σ with $\|\varphi\|_{C^\alpha} \leq 1$, $\|\sigma\|_{C^\alpha} \leq 1$. These limits can be expressed by

$$V\sigma(x) = \int_{y \in \Gamma \setminus \{x\}} E(x, y) \sigma(y) \, ds_y, \quad x \in \Gamma \quad (20)$$

$$K\varphi(x) = \text{p.v.} \int_{y \in \Gamma \setminus \{x\}} (T_y E(x, y))^T \varphi(y) \, ds_y, \quad x \in \Gamma \quad (21)$$

$$K'\sigma(x) = \text{p.v.} \int_{y \in \Gamma \setminus \{x\}} (T_x E(x, y)) \sigma(y) \, ds_y, \quad x \in \Gamma \quad (22)$$

If, in addition, φ is Hölder continuously differentiable, then the limit in (19) exists uniformly with respect to all $x \in \Gamma$ and all φ with $\|\varphi\|_{C^{1+\alpha}} \leq 1$. This limit can be expressed by

$$\begin{aligned} D\varphi(x) &= - \lim_{\Omega z \rightarrow x \in \Gamma} T_z \int_\Gamma (T_y E(z, y))^T [\varphi(y) - \varphi(x)] \, ds_y \\ &= -\text{p.v.} \int_\Gamma T_x (T_y E(x, y))^T [\varphi(y) - \varphi(x)] \, ds_y \end{aligned} \quad (23)$$

Here and in the sequel, we denote by

$$C^{m+\alpha}(\Gamma) := \{v \in C^m(\Gamma) \mid \|v\|_{C^{m+\alpha}(\Gamma)} < \infty\}$$

the Hölder m -continuously differentiable function space equipped with the norm

$$\begin{aligned} \|v\|_{C^{m+\alpha}(\Gamma)} &= \sum_{|\beta| \leq m} \sup_{x \in \Gamma} |\partial^\beta v(x)| \\ &\quad + \sum_{|\beta|=m} \sup_{x, y \in \Gamma, x \neq y} \frac{|\partial^\beta v(x) - \partial^\beta v(y)|}{|x - y|^\alpha} \end{aligned}$$

for $m \in \mathbb{N}_0$, the set of nonnegative integers, and $0 < \alpha < 1$, where ∂^β denotes the covariant derivatives

$$\partial^\beta := \partial_1^{\beta_1} \cdots \partial_{n-1}^{\beta_{n-1}}$$

on the $(n-1)$ -dimensional boundary surface Γ and $\beta \in \mathbb{N}_0^{n-1}$ denotes the multi-index with $|\beta| = \beta_1 + \cdots + \beta_{n-1}$ (see Millman and Parkor, 1977).

The operator V has weakly singular kernel, K and K' have Cauchy-singular kernels, whereas D has a nonintegrable kernel with singularity of the order $O(|x - y|^{-\alpha})$. The integral in (20) is a weakly singular improper integral, whereas the integrals in (21) and (22) as well as in (23) are to be defined as *Cauchy principal value integrals*, for example,

$$\begin{aligned} \text{p.v.} \int_{y \in \Gamma \setminus \{x\}} (T_y E(x, y))^T \varphi(y) \, ds_y \\ = \lim_{\varepsilon \rightarrow 0} \int_{|y-x| \geq \varepsilon \wedge y \in \Gamma} (T_y E(x, y))^T \varphi(y) \, ds_y \end{aligned}$$

If we apply the representation formula to any constant vector field a , which represents a rigid displacement, then we obtain

$$a = - \int_\Gamma (T_y E(z, y))^T a \, ds_y \quad \text{for } z \in \Omega$$

which yields for $z \in \Omega$ in the neighborhood of Γ

$$T_z \int_\Gamma (T_y E(z, y))^T \varphi(y) \, ds_y = 0$$

Hence,

$$D\varphi(x) = \lim_{\Omega z \rightarrow x \in \Gamma} T_z \int_\Gamma (T_y E(z, y))^T (\varphi(y) - \varphi(x)) \, ds_y$$

from which (23) results. For a proof of Lemma 1, see Mikhlin (1970). In fact, for $\Gamma \in C^2$ and $0 < \alpha < 1$, a fixed constant, these boundary integral operators define continuous mappings between the following spaces:

$$\begin{aligned} V: C^\alpha(\Gamma) &\rightarrow C^{1+\alpha}(\Gamma) \\ K, K': C^\alpha(\Gamma) &\rightarrow C^\alpha(\Gamma), \quad C^{1+\alpha}(\Gamma) \rightarrow C^{1+\alpha}(\Gamma) \\ D: C^{1+\alpha}(\Gamma) &\rightarrow C^\alpha(\Gamma) \end{aligned}$$

(see Mikhlin (1970) and Mikhlin and Prüssdorf (1986)).

We remark that from the overdetermined system of BIEs (14) and (15) for any solution of (8), the Cauchy data φ, σ

on Γ are reproduced by the matrix operator on the right-hand side of (14) and (15), namely,

$$C_\Omega := \begin{pmatrix} \tfrac{1}{2}I - K & V \\ D & \tfrac{1}{2}I + K' \end{pmatrix}$$

This suggests that this matrix operator C_Ω is a projector. Indeed, this is the case and the matrix operator C_Ω is the so-called *Calderon projector*. From the mapping properties of the corresponding boundary integral operators, we have the following basic result.

Theorem 1. Let $\Gamma \in C^2$. Then, C_Ω maps $C^\alpha(\Gamma) \times C^{1+\alpha}(\Gamma)$ into itself continuously and C_Ω is a projector in the sense that

$$C_\Omega^2 = C_\Omega$$

As a consequence of Theorem 1, we have the following identities:

$$\begin{aligned} VD = \tfrac{1}{2}I - K^2, \quad DV = \tfrac{1}{2}I - K'^2 \\ KV = VK', \quad DK = K'D \end{aligned} \quad (24)$$

for the four boundary integral operators. We remark that these relations are extremely valuable from both theoretical and computational points of view. In particular, V and D are pseudoinverses to each other and may serve as preconditioners in the corresponding variational formulations (see Steinbach and Wendland, 1998). The Calderon projector leads in a direct manner to the basic BIEs for the BVPs under consideration.

Now for the exterior problems, we begin with the representation (12) and obtain, in the same manner, the system of BIEs on Γ :

$$\varphi = (\tfrac{1}{2}I + K)\varphi - V\sigma + \omega \quad (25)$$

$$\sigma = -D\varphi + (\tfrac{1}{2}I - K')\sigma \quad (26)$$

for the Cauchy data φ, σ of the solution $u \in \Omega^c$. Here we have used the fact that the four basic boundary integral operators V, K, K' , and D are related to the limits of the boundary potentials from the exterior domain Ω^c , similar to (16) to (19), namely,

$$V\sigma(x) = \lim_{\Omega^c z \rightarrow x \in \Gamma} V\sigma(z) \quad (27)$$

$$K\varphi(x) = \lim_{\Omega^c z \rightarrow x \in \Gamma} W\varphi(z) - \tfrac{1}{2}\varphi(x) \quad (28)$$

$$K'\sigma(x) = \lim_{\Omega^c z \rightarrow x \in \Gamma} T_z V\sigma(z) + \tfrac{1}{2}\sigma(x) \quad (29)$$

$$D\varphi(x) = - \lim_{\Omega^c z \rightarrow x \in \Gamma} T_z W\varphi(z) \quad (30)$$

for $x \in \Gamma$. Note that in (28) and (29), the signs at $(1/2)\varphi(x)$ and $(1/2)\sigma(x)$ are different from those in (17) and (18) respectively and that the latter provide us the so-called *jump relations* of the boundary potentials across the boundary.

For any solution u of (11) in Ω^c with $\omega = 0$, we may define similarly the *Calderon projector* C_{Ω^c} for the exterior domain Ω^c . There holds the relation

$$C_{\Omega^c} = I - C_{\Omega}$$

where I stands for the identity matrix operator and Theorem 1 remains valid for C_{Ω^c} as well.

2.3 Boundary integral equations

The interior displacement problem consists of the Lamé equations (8) in Ω together with the prescribed Dirichlet condition (2), namely,

$$u|_{\Gamma} = \varphi \quad \text{on } \Gamma \quad (31)$$

where we have tacitly replaced ϕ by $\varphi = \phi - u_p|_{\Gamma}$. The missing Cauchy datum on Γ is the boundary traction $\sigma = T u_p$. For the solution of the interior displacement problem, we may solve either (14), the Fredholm BIE of the first kind

$$V\sigma = \left(\frac{1}{2}I + K\right)\varphi \quad \text{on } \Gamma \quad (32)$$

or (15), the Cauchy-singular integral equation (CSIE) of the second kind

$$\left(\frac{1}{2}I - K'\right)\sigma = D\varphi \quad \text{on } \Gamma \quad (33)$$

Both are BIEs for the unknown σ .

For $n = 2$, the first-kind integral equation (32) may have eigensolutions for special choices of Γ . To circumvent this difficulty, we can modify (32) by including rigid motions (10). More precisely, we consider the system

$$V\sigma - \omega = \left(\frac{1}{2}I + K\right)\varphi \quad \text{on } \Gamma \quad \text{and} \quad \int_{\Gamma} \sigma \, ds = 0 \quad (34)$$

together with

$$\begin{aligned} \int_{\Gamma} (-\sigma_1 x_2 + \sigma_2 x_1) \, ds &= 0 \quad \text{for } n = 2 \\ \text{or} \quad \int_{\Gamma} (\sigma \times (x_1, x_2, x_3)^{\top}) \, ds &= 0 \quad \text{for } n = 3 \end{aligned} \quad (35)$$

where σ and ω (see 10) are to be determined. As was shown in Hsiao and Wendland (1985), the rotation b (b for $n = 2$) in (10) can be prescribed as $b = 0$ ($b = 0$

for $n = 2$); in this case, the side conditions (35) will not be needed. In fact, for $n = 3$, many more choices in ω can be made (see Hsiao and Wendland (1985) for the details). Now, given $\varphi \in C^{1+\alpha}(\Gamma)$, the modified system (34), (35) is always uniquely solvable for $\sigma \in C^{\alpha}(\Gamma)$ in the Hölder space (for $n = 2$, the analysis is based on, for example, Muskhelishvili (1953), Fichera (1961), and Hsiao and MacCamy (1973)).

For the special CSIE of the second kind (33), Mikhlin (1962) showed that the Fredholm alternative – originally designed for compact operators – remains valid here. Therefore, (33) admits a unique classical solution $\sigma \in C^{\alpha}(\Gamma)$, provided $\varphi \in C^{1+\alpha}(\Gamma)$, $0 < \alpha < 1$; see Kupradze *et al.* (1979) and Mikhlin and Prössdorf (1986).

The interior traction problem consists of the Lamé system (8) in Ω together with the prescribed Neumann condition (3):

$$T u|_{\Gamma} = \sigma \quad \text{on } \Gamma \quad (36)$$

where again we have replaced ψ by $\sigma = \psi - T u_p|_{\Gamma}$. The missing Cauchy datum is now $\varphi = u|_{\Gamma}$ in the overdetermined system (14) and (15). Again, we can solve for φ from either (14)

$$\left(\frac{1}{2}I + K\right)\varphi = V\sigma \quad \text{on } \Gamma \quad (37)$$

or from (15)

$$D\varphi = \left(\frac{1}{2}I - K'\right)\sigma \quad \text{on } \Gamma \quad (38)$$

with given σ . As is well known, similar to the Neumann problem for the Laplacian, the given traction σ needs to satisfy equilibrium conditions for a solution of the interior traction problem to exist. These can be obtained from the second Green formula, which is known as the *Betti formula* in elasticity. The equilibrium condition reads

$$\int_{\Gamma} \omega \cdot \sigma \, ds = 0 \quad (39)$$

for all the rigid motions ω given by (10). Or, equivalently, from the definition of ω in (10), this means that σ should satisfy

$$\begin{aligned} a \cdot \int_{\Gamma} \sigma \, ds + b \int_{\Gamma} (-\sigma_1 x_2 + \sigma_2 x_1) \, ds &= 0 \quad \text{for } n = 2 \\ \text{or} \quad a \cdot \int_{\Gamma} \sigma \, ds + b \cdot \int_{\Gamma} (x_1, x_2, x_3)^{\top} \times \sigma \, ds &= 0 \quad \text{for } n = 3 \end{aligned}$$

for arbitrary $b \in \mathbb{R}$, a and $b \in \mathbb{R}^3$. This condition also turns out to be sufficient for the existence of φ in the classical Hölder-function spaces. If $\sigma \in C^{\alpha}(\Gamma)$ with $0 < \alpha < 1$ is given satisfying (39), then the right-hand side $V\sigma$ in (37)

automatically satisfies the orthogonality conditions from Fredholm's alternative, and the CSIE (37) admits a solution $\varphi \in C^{1+\alpha}(\Gamma)$. However, the solution is unique only up to all rigid motions ω , which are eigensolutions. For further details, see Kupradze *et al.* (1979).

The hypersingular integral equation of the first kind (38) also has eigensolutions, which again are given by all rigid motions (10). However, the classical Fredholm alternative also holds for (38); and the right-hand side $[(1/2)I - K']\sigma$ in (38) satisfies the corresponding orthogonality conditions, provided $\sigma \in C^{\alpha}(\Gamma)$ satisfies the equilibrium conditions (39). In both cases, the integral equations, together with appropriate side conditions, can be modified so that the resulting equations are uniquely solvable.

For the exterior displacement problem, u satisfies the Lamé system (11) in Ω^c , the Dirichlet condition (31), and the decay condition (9) at infinity with given total surface forces Σ . For simplicity, we consider here only the case in which the rigid motions ω are unknown. For other cases, we refer the interested reader to the work in Hsiao and Wendland (1985). For the present simple situation, both the tractions $\sigma = T u|_{\Gamma}$ and the rigid motions ω are now the unknowns. We may solve the problem by using either (25) or (26). However, for both equations, in addition to the given total surface force as in (13), we need to prescribe additional normalization conditions, for example, the total moment condition due to boundary traction as in (35), in order to have the uniqueness of the solution. This yields the following modified BIE of the first kind from (25):

$$V\sigma - \omega = -\left(\frac{1}{2}I - K\right)\varphi \quad \text{on } \Gamma \quad (40)$$

$$\int_{\Gamma} \sigma \, ds = \Sigma \quad (41)$$

together with the additional normalization conditions

$$\begin{aligned} \int_{\Gamma} (-\sigma_1 x_2 + \sigma_2 x_1) \, ds &= 0 \quad \text{for } n = 2 \\ \text{or} \quad \int_{\Gamma} (\sigma \times (x_1, x_2, x_3)^{\top}) \, ds &= 0 \quad \text{for } n = 3 \end{aligned} \quad (42)$$

where $\varphi \in C^{1+\alpha}(\Gamma)$ and $\Sigma \in \mathbb{R}^n$, $n = 2, 3$ are given, but σ and ω are to be determined. It can be shown that (40) together with (41) and (42) always has a unique solution $\sigma \in C^{\alpha}(\Gamma)$ and ω , the rigid motion in the form of (10).

On the other hand, from (26) we have the singular integral equation of the second kind

$$\left(\frac{1}{2}I + K'\right)\sigma = -D\varphi \quad \text{on } \Gamma \quad (43)$$

with the additional equations (41) and (42). Note that the operator $(1/2)I + K'$ is adjoint to $(1/2)I + K$. Owing to

Mikhlin (1962), for these special operators, the classical Fredholm alternative is still valid in the space $C^{\alpha}(\Gamma)$. Since $((1/2)I + K)\omega = 0$ on Γ for all rigid motions ω , the corresponding homogeneous equation (43) also has a $3(n-1)$ -dimensional eigenspace. Moreover, $D\omega = 0$ for all rigid motions; hence, the right-hand side of (43) always satisfies the orthogonality conditions for any given $\varphi \in C^{1+\alpha}(\Gamma)$. This implies that equation (43) always admits a solution $\sigma \in C^{\alpha}(\Gamma)$. With the prescribed total force (41) and total moment due to traction (42), additionally, these algebraic equations determine σ uniquely.

Finally, for the exterior traction problem, the Neumann datum is given by (36) and so are the total forces (41) and total moment such as (42), although here the total moment may or may not be equal to zero, depending on the given traction σ on Γ . The rigid motion ω is now an additional parameter, which can be prescribed arbitrarily according to the special situation. Often, $\omega = 0$ is chosen in the representation (12) as well as in the decay condition (9) for the exterior traction problem. Then, from (25) we have the Cauchy-singular BIE for $\varphi = u|_{\Gamma}$:

$$\left(\frac{1}{2}I - K\right)\varphi = -V\sigma + \omega \quad \text{on } \Gamma \quad (44)$$

As is well known, for any given $\sigma \in C^{\alpha}(\Gamma)$ and given ω , the equation (44) is always uniquely solvable for $\varphi \in C^{1+\alpha}(\Gamma)$; we refer the reader to Kupradze (1965) for details.

We may also solve the exterior traction problem by using (26). This leads to the hypersingular BIE of the first kind,

$$D\varphi = -\left(\frac{1}{2}I + K'\right)\sigma \quad \text{on } \Gamma \quad (45)$$

It is easily seen that rigid motions ω on Γ are eigensolutions of (45). Therefore, in order to guarantee unique solvability of the BIE, we modify (45) by including additional restrictions and adding more unknowns, for example,

$$D\varphi_0(x) + \sum_{\ell=1}^{3(n-1)} \alpha_{\ell} m_{\ell}(x) = -\left(\frac{1}{2}I + K'\right)\sigma(x) \quad \text{for } x \in \Gamma \quad \text{and} \quad \int_{\Gamma} m_{\ell}(y) \cdot \varphi_0(y) \, ds = 0, \quad \ell = 1, \dots, 3(n-1) \quad (46)$$

where $m_{\ell}(x)$ denote the basis vectors for the rigid motions ω defined by (10). Here the added unknown Lagrangian multipliers α_{ℓ} are introduced in order to have the same number of unknowns as that of the equations. It is easy to see that for the Neumann problem as the exterior traction problem, they all vanish because of the special form of the right-hand side of (46). This system is always uniquely solvable, and for any given $\sigma \in C^{\alpha}(\Gamma)$, we find exactly one

$\phi_0 \in C^{1+\alpha}(\Gamma)$. Once ϕ_0 is known, the displacement field $u(x)$ in Ω^c is given by the representation

$$u(x) = -V\sigma(x) + W\phi_0(x) \quad \text{for } x \in \Omega^c \quad (47)$$

Note that the actual boundary values of $u|_\Gamma$ may differ from ϕ_0 by a rigid motion and can be expressed via (47) in the form

$$u(x)|_\Gamma = (\frac{1}{2}I + K)\phi_0(x) - V\sigma(x) \quad \text{for } x \in \Gamma$$

In summary, we see that for each of these fundamental problems, the solution may be obtained by solving the BIE of either the first kind or the second kind. Moreover, the unique solvability of these BIEs can be accomplished by including appropriate normalization conditions based on the rigid motions.

2.4 Boundary elements and numerical schemes

We observe that all the BIEs derived previously can be formally written in the form

$$\begin{aligned} Av + B\alpha &= q \\ \Lambda v &= c \end{aligned} \quad (48)$$

Here, v and α denote the unknown Cauchy datum and real vectors, while q and c are the given data on the boundary and constant constraint vectors respectively; A is a boundary integral operator, B is a matrix of functions, and Λ consists of appropriate functionals. We now discuss the numerical solutions of (48) by the boundary elements.

For the numerical treatment of BIEs (48), one needs a discretization of the boundary charge functions v . In the BEM, trial functions are chosen as finite elements on the boundary Γ , the boundary elements. In general, the boundary Γ is assumed to be given by local parametric representations such that a family of regular partitions in the parametric domains is mapped onto corresponding partitions on Γ . On the partitions of the parametric domain, we use a regular family of finite elements as introduced in Babuška and Aziz (1977). The parametric representation of Γ then transfers the finite elements onto the boundary Γ defining a family of boundary elements. However, in practice, or if the local parametric representations of Γ are not known explicitly, an approximation Γ_h of the boundary surface Γ is often used. The approximated boundary Γ_h is composed of piecewise polynomials associated with a family of triangulations $\{\tau_i\}_{i=1}^N$ of Γ , where $\Gamma = \bigcup_{i=1}^N \tau_i$ and $h := \max_{i=1, \dots, N} L(\tau_i)$. The degree of the piecewise polynomials is chosen in correspondence to the order of the Sobolev spaces for the

solutions of the BIEs. In this connection, we refer the reader to the work by Nedelec (1976) and also Dautry and Lions (1990).

There are three main numerical schemes for treating the BIEs. These are *collocation*, *Galerkin's method*, and the *least squares method*. For the BIE (48), these methods can be described as follows. Let S_h denote a family of finite-dimensional subspaces of the solution space for v based on the triangulation. Let $\mu_1(x), \dots, \mu_N(x)$ be a basis of S_h for fixed h . We seek an approximate solution pair $\{v_h, \alpha_h\}$ in the form

$$v_h(x) = \sum_{j=1}^N \gamma_j \mu_j(x), \quad \alpha_h \in \mathbb{R}^{3(n-1)} \quad (49)$$

with unknown coefficients γ_j and vector α_h to be determined by following a system of algebraic equations generated by these methods:

(1) *The collocation method.* A suitable set of collocation points $x_k \subset \Gamma$, $k = 1, \dots, N$ is chosen and the approximate solution (49) is required to satisfy the collocation equations

$$\sum_{j=1}^N \gamma_j A \mu_j(x_k) + B(x_k) \alpha_h = q(x_k), \quad k = 1, \dots, N$$

$$\sum_{j=1}^N \gamma_j \Lambda \mu_j = c \quad (50)$$

(2) *The Galerkin method.* The approximate solution (49) is required to satisfy the Galerkin equations

$$\sum_{j=1}^N \gamma_j (A \mu_j, \mu_k) + (B \alpha_h, \mu_k) = (q, \mu_k), \quad k = 1, \dots, N$$

$$\sum_{j=1}^N \gamma_j \Lambda \mu_j = c \quad (51)$$

where the brackets (\cdot, \cdot) denote the $L^2(\Gamma)$ -inner product,

$$(\sigma, \varphi) := \int_\Gamma \sigma(y) \cdot \overline{\varphi(y)} dy$$

where $\overline{\varphi}$ denotes the complex conjugate of φ . For the equations (46), for instance, the equations (51) correspond to the mixed formulation of saddlepoint problems. As will be seen, the Galerkin equation (51) represents, in some sense, the discrete weak formulation of (48) in the finite-dimensional subspace S_h ; the precise definition will be made clear in the next two sections.

(3) *The least squares method.* The approximate solution (49) is determined from the condition

$$\left\| \sum_{j=1}^N \gamma_j A \mu_j + B \alpha_h - q \right\|_{L^2(\Gamma)}^2 + \left\| \sum_{j=1}^N \gamma_j A \mu_j - c \right\|^2 = \min!$$

or, equivalently, from the linear system of algebraic equations

$$\begin{aligned} \sum_{j=1}^N \gamma_j (A \mu_j, A \mu_k) + (B \alpha_h, A \mu_k) + \sum_{j=1}^N \gamma_j \Lambda \mu_j \cdot \Lambda \mu_k \\ = (q, A \mu_k) + c \cdot \Lambda \mu_k, \quad k = 1, \dots, N \\ \sum_{j=1}^N \gamma_j (\Lambda \mu_j, B^T) + (B \alpha_h, B^T) = (q, B^T) \end{aligned}$$

This completes the tutorial part of our introduction to the BEM. Needless to say, the solvability of these linear systems as well as the error estimates for the approximate solutions depend heavily on the properties of the boundary integral operators involved, and all the formal formulations can be made precise after we introduce appropriate mathematical tools. We will return to these discussions later.

3 VARIATIONAL FORMULATIONS

To discuss the Galerkin-BEM, we need the concept of weak or variational solutions of the BIEs in the same manner as that of the PDEs. This section will be devoted to the above topic by using a simple Dirichlet problem for the Laplacian (instead the Lamé) operator to illustrate all the essential features in the variational formulations, motivating from the weak solution of the PDEs to that of the BIEs. Our analysis is general enough and can immediately be applied to elasticity without any severe modifications (see Hsiao and Wendland, 1985) (see Chapter 4, this Volume).

3.1 Weak solutions

To introduce the concept of weak solutions for the BIEs, let us begin with the variational formulation of the simplest elliptic BVP, the Dirichlet problem for Poisson's equation,

$$-\Delta u = f \quad \text{in } \Omega \subset \mathbb{R}^3 \quad \text{and } u|_\Gamma = 0 \quad \text{on } \Gamma \quad (52)$$

where f is a given function satisfying certain regularity conditions to be specified. Multiplying both sides of the

equation $-\Delta u = f$ by a smooth function \bar{v} and integrating by parts, we obtain

$$\int_\Omega \nabla u(x) \cdot \nabla \bar{v}(x) dx = \int_\Omega f(x) \bar{v}(x) dx \quad (53)$$

provided $v|_\Gamma = 0$. The integral identity (53) suggests that one may seek a solution of the Dirichlet problem (52) from (53) when the classical solution does not exist, as long as the integrals in (53) are meaningful. Indeed, this leads to the concept of the variational formulations of BVPs and the weak solutions for partial differential equations. To make it more precise, let us introduce the function space

$$H^1(\Omega) = \{v \in L^2(\Omega) | \nabla v \in L^2(\Omega)\}$$

equipped with the norm

$$\|v\|_{H^1(\Omega)} := \left(\int_\Omega |v(x)|^2 + |\nabla v(x)|^2 dx \right)^{1/2}$$

the Sobolev space of the first order, which is a Hilbert space with the inner product

$$(v, w)_{H^1(\Omega)} := \int_\Omega \{v(x) \overline{w(x)} + \nabla v(x) \cdot \nabla \overline{w(x)}\} dx$$

We denote by

$$H_0^1(\Omega) = \{v \in H^1(\Omega) | v|_\Gamma = 0\}$$

the subspace of $H^1(\Omega)$ with homogeneous Dirichlet boundary conditions. We denote the dual space of $H^1(\Omega)$ by $\tilde{H}^{-1}(\Omega)$, while we denote the dual space of $H_0^1(\Omega)$ by $H^{-1}(\Omega)$. The corresponding norms are defined by

$$\|f\|_{\tilde{H}^{-1}(\Omega)} = \sup_{v \in H_0^1(\Omega)} \frac{|(f, v)|}{\|v\|_{H^1(\Omega)}}$$

and

$$\|f\|_{H^{-1}(\Omega)} = \sup_{v \in H^1(\Omega)} \frac{|(f, v)|}{\|v\|_{H^1(\Omega)}}$$

$\tilde{H}^{-1}(\Omega)$ and $H^{-1}(\Omega)$ contain all the bounded linear functionals on $H^1(\Omega)$ or $H_0^1(\Omega)$ respectively. We adopt the notion (\cdot, \cdot) for the L^2 -duality pairings between $\tilde{H}^{-1}(\Omega)$ and $H^1(\Omega)$, and $H^{-1}(\Omega)$ and $H_0^1(\Omega)$ respectively; if $f \in L^2(\Omega)$, we have simply $(f, v) = (f, v)_{L^2(\Omega)}$ for all $v \in H^1(\Omega)$. With these function spaces available, we may now give the precise definition of a weak solution of (52). Given $f \in H^{-1}(\Omega)$, $u \in H_0^1(\Omega)$ is said to be a *weak solution* of (52), if it satisfies

$$a_\Omega(u, v) = \ell_f(v) \quad \text{for all } v \in H_0^1(\Omega) \quad (54)$$

Here, in the formulation, we have introduced the bilinear form

$$a_\Omega(u, v) := \int_\Omega \nabla u \cdot \nabla v \, dx$$

while for fixed $f \in H^{-1}(\Omega)$ by $\ell_f(v)$, we denote the anti-linear (i.e. conjugate-linear) functional $\ell_f: H_0^1(\Omega) \ni v \mapsto \ell_f(v) \in \mathbb{R}$ defined by $\ell_f(v) := (f, v)$. By definition, the weak solution of (52) is a function from $H_0^1(\Omega)$, which thus does not need to have derivatives of the second order in Ω (it has only generalized derivatives of the first order, in general). The concept of weak solutions of the BVP (52) is thus substantially more general than the concept of classical solutions of the problem.

It is well known that (54) has a unique solution. In fact, the existence and uniqueness of the solution of (54) can be established by the celebrated Lax–Milgram Lemma (see Lax and Milgram, 1954). In the present example, since the bilinear form $a(\cdot, \cdot)$ is symmetric, that is, $a(u, v) = \overline{a(v, u)}$ for all $u, v \in H_0^1(\Omega)$, it can be shown that the solution u of (54) is the only element in $H_0^1(\Omega)$ that minimizes in the real Hilbert space $H_0^1(\Omega)$ the following quadratic functional:

$$J(v) := \frac{1}{2} a_\Omega(v, v) - \ell_f(v) \quad (55)$$

This is the well-known *Dirichlet principle*, and the quadratic functional in (55) is called the *energy functional*. Consequently, the solution space $H_0^1(\Omega)$ is often referred to as the energy space of the Dirichlet problem for equation (52). Dirichlet's principle provides an alternative means of approximating the solution of (54) by minimizing $J(\cdot)$ in finite element subspaces of $H_0^1(\Omega)$. This is known as the *Rayleigh–Ritz method* and leads to the same linear system of algebraic equations as obtained by the *Galerkin method* in the present situation. Although both methods lead to the same results, the basic ideas of these methods are different.

In general, the reinterpretation of the Dirichlet problem (52) in the context of distribution theory, here in the form of (54), is often referred to as the *weak formulation* of the problem, whereas a minimization problem in the form of (55) is referred to as the *variational formulation* of (52), and the equation (54) is referred to as the *variational equation*. However, these terminologies are not universal. Throughout the sequel, these terms will be used interchangeably without distinction.

For the BIEs, without loss of generality, we will first modify the problem (52) slightly by considering the homogeneous equation but nonhomogeneous Dirichlet boundary condition, namely,

$$\begin{aligned} \Delta u &= 0 \quad \text{in } \Omega \subset \mathbb{R}^3 \\ u|_\Gamma &= \varphi \quad \text{on } \Gamma \end{aligned} \quad (56)$$

where $\varphi = \bar{\varphi}|_\Gamma$ for some given function $\bar{\varphi} \in H^1(\Omega)$. Here, φ is the given Cauchy datum and the missing one is the normal derivative of u , that is,

$$\sigma := \nabla u \cdot \mathbf{n}|_\Gamma = \frac{\partial u}{\partial n}$$

in terms of our terminology in Section 2. Again, one may show that there is a unique weak solution u , which is now in $H^1(\Omega)$ such that $u - \bar{\varphi} \in H_0^1(\Omega)$ and

$$a_\Omega(u, v) = 0 \quad \text{for all } v \in H_0^1(\Omega)$$

To discuss the weak solution of the BIEs, we again introduce the *simple- and double-layer potentials*

$$V\lambda(x) := \int_\Gamma E(x, y)\lambda(y) \, dy$$

and

$$W\mu(x) := \int_\Gamma \frac{\partial}{\partial n_y} E(x, y)\mu(y) \, dy \quad \text{for } x \in \Omega \quad (57)$$

for density functions λ and μ , where E is the fundamental solution for the Laplacian in \mathbb{R}^3 given by

$$E(x, y) := \frac{1}{4\pi} \frac{1}{|x - y|}$$

(see e.g. equation 5). In the *indirect approach* based on the layer-ansatz, for smooth density functions λ and μ , we may seek a solution of (56) in the form of a simple-layer potential $u(x) = V\lambda(x)$, or in the form of a double-layer potential $u(x) = -W\mu(x)$ for $x \in \Omega$. The former then leads to an *integral equation of the first kind*,

$$V\lambda = \varphi \quad \text{on } \Gamma \quad (58)$$

while the latter leads to an *integral equation of the second kind*

$$\left(\frac{1}{2}I - K\right)\mu = \varphi \quad \text{on } \Gamma \quad (59)$$

Here, V is the simple-layer boundary integral operator on the boundary

$$V\lambda(x) := \int_\Gamma E(x, y)\lambda(y) \, dy \quad \text{for } x \in \Gamma \quad (60)$$

where we have kept the same notation V for the boundary integral operator, while the boundary integral operator K , defined by

$$K\mu(x) := \int_{\Gamma \setminus \{x\}} \frac{\partial}{\partial n_y} E(x, y)\mu(y) \, dy \quad \text{for } x \in \Gamma$$

is now the boundary double-layer potential operator. The operator V has a weakly singular kernel. But in contrast to the situation in elasticity, for the Laplacian both K and its transposed

$$K'\lambda(x) := \int_{\Gamma \setminus \{x\}} \frac{\partial}{\partial n_x} E(x, y)\lambda(y) \, dy \quad \text{for } x \in \Gamma$$

also have weakly singular kernels, provided Γ is smooth enough (e.g. $\Gamma \in C^2$); then these operators are also well defined on the classical Hölder spaces. In order to define weak solutions of the BIEs (58) and (59) in the same manner as for the PDEs, and to be able to extend our approach to more general Γ as to Lipschitz boundaries, we need the density functions in appropriate Sobolev spaces. In other words, we need the right boundary energy spaces for the BIEs. These should be in some sense closely related to the 'traces' of the energy spaces from the BVPs with partial differential equations. Indeed, this can be seen as follows.

First, observe that for a given smooth density function λ (say, in $C^\infty(\Gamma)$), the potential

$$u(x) = V\lambda(x) := \int_\Gamma E(x, y)\lambda(y) \, dy \quad \text{for } x \in \mathbb{R}^3 \setminus \Gamma \quad (61)$$

satisfies $\Delta u = 0$ in Ω as well as in Ω^c . Moreover, we have the jump relation

$$\lambda = \frac{\partial u^-}{\partial n} - \frac{\partial u^+}{\partial n} =: \left[\frac{\partial u}{\partial n} \right]_\Gamma \quad (62)$$

where u^\pm denotes the limits of the function u on Γ from Ω^\pm and Ω respectively. From Green's formula, we obtain the relation

$$\int_\Gamma \lambda(y) V\lambda(y) \, dy = \int_\Gamma \left[\frac{\partial u}{\partial n} \right]_\Gamma u|_\Gamma \, dy = \int_{\Omega \cup \Omega^c} |\nabla u(x)|^2 \, dx \quad (63)$$

On the other hand, for a given smooth moment function μ (say, in $C^{1+\alpha}(\Gamma)$), set

$$\begin{aligned} u(x) &= -W\mu(x) := - \int_\Gamma \frac{\partial}{\partial n_y} E(x, y)\mu(y) \, dy \\ &\quad \text{for } x \in \mathbb{R}^3 \setminus \Gamma \end{aligned} \quad (64)$$

Similarly, u satisfies $\Delta u = 0$ in Ω as well as in Ω^c and the corresponding jump relation reads

$$\mu = u^- - u^+ =: [u]_\Gamma$$

An application of the Green formula then yields the relation

$$\begin{aligned} \int_\Gamma D\mu(y) \left(\frac{1}{2}I - K\right)\mu(y) \, dy &= \int_\Gamma \frac{\partial u^-}{\partial n} u^- \, dy \\ &= \int_\Omega |\nabla u(x)|^2 \, dx \end{aligned} \quad (65)$$

where we have denoted by D the *hypersingular boundary integral operator* for (56) defined by

$$D\mu(x) := - \frac{\partial}{\partial n_x} \int_\Gamma \frac{\partial}{\partial n_y} E(x, y)\mu(y) \, dy \quad \text{for } x \in \Gamma$$

In view of the identities (63) and (65), it is not difficult to see how the exact weak formulations of the BIEs of the first kind (58) and of the second kind (59) should be formulated in the Sobolev spaces. Moreover, these relations also suggest appropriate energy boundary spaces for the corresponding bilinear forms for the boundary integral operators under consideration. To make all this precise, we need the concept of the boundary value of the function $u \in H^1(\Omega)$ on Γ .

Let $L^2(\Gamma)$ be the space of square integrable functions φ on Γ , equipped with the norm

$$\|\varphi\|_{L^2(\Gamma)} = \left(\int_\Gamma |\varphi(y)|^2 \, dy \right)^{1/2}$$

Then, one can show that there is a constant c such that

$$\|\gamma_0 u\|_{L^2(\Gamma)} \leq c \|u\|_{H^1(\Omega)} \quad \text{for all } u \in C^1(\bar{\Omega})$$

where $\gamma_0 u = u|_\Gamma$ denotes the boundary value of u on Γ . By continuity, the mapping γ_0 defined on $C^1(\bar{\Omega})$ can be extended to a mapping, still called γ_0 , from $H^1(\Omega)$ into $L^2(\Gamma)$. By the extension, $\gamma_0 u$ is called the boundary value of u on Γ . It can be shown that

$$\text{Ker}(\gamma_0) := \{u \in H^1(\Omega) \mid \gamma_0 u = 0\} = H_0^1(\Omega)$$

and the range of γ_0 is a proper and dense subspace of $L^2(\Gamma)$, called $H^{1/2}(\Gamma)$. For $\varphi \in H^{1/2}(\Gamma)$, we define the norm

$$\|\varphi\|_{H^{1/2}(\Gamma)} := \inf_{\substack{u \in H^1(\Omega) \\ \gamma_0 u = \varphi}} \|\varphi\|_{H^1(\Omega)}$$

and $H^{1/2}(\Gamma)$ is a Hilbert space. We denote by $H^{-1/2}(\Gamma)$ the corresponding dual space of $H^{1/2}(\Gamma)$, equipped with the norm

$$\|\lambda\|_{H^{-1/2}(\Gamma)} := \sup_{\substack{\varphi \in H^{1/2}(\Gamma) \\ \|\varphi\|_{H^{1/2}(\Gamma)} = 1}} |(\lambda, \mu)|$$

where (\cdot, \cdot) now denotes the $L^2(\Gamma)$ -duality pairing between the spaces $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$ on Γ . For $\lambda \in L^2(\Gamma)$, the duality (λ, μ) is simply identified with $(\lambda, \mu)_{L^2(\Gamma)}$.

We remark that the Cauchy data $u|_\Gamma$ and $\partial u/\partial n$ of the weak solution $u \in H^1(\Omega)$ of (56) are in $H^{1/2}(\Gamma)$ and its dual $H^{-1/2}(\Gamma)$ respectively, as will be seen from the generalized first Green theorem to be stated later. We may define from (63) and (65) the boundary sesquilinear forms

$$\begin{aligned} a_{1\Gamma}(\chi, \lambda) &:= (\chi, V\lambda) \quad \text{for all } \chi, \lambda \in H^{-1/2}(\Gamma) \quad (66) \\ a_{2\Gamma}(\nu, \mu) &:= (D\nu, (\tfrac{1}{2}I - K)\mu) \quad \text{for all } \nu, \mu \in H^{1/2}(\Gamma) \quad (67) \end{aligned}$$

Note that both the sesquilinear forms $a_{1\Gamma}$ and $a_{2\Gamma}$ are symmetric (or Hermitian) since V and D are both self-adjoint and, moreover, $DK = K'D$ because of the Calderon projection property. Then the weak formulation of the BIE of the first kind (58) reads as follows:

Definition 1. Given $\varphi \in H^{1/2}(\Gamma)$, find a function $\lambda \in H^{-1/2}(\Gamma)$ such that

$$a_{1\Gamma}(\chi, \lambda) = \ell_\varphi(\chi) \quad \text{for all } \chi \in H^{-1/2}(\Gamma) \quad (68)$$

where ℓ_φ is the linear functional on $H^{-1/2}(\Gamma)$, defined by $\ell_\varphi(\chi) := (\chi, \varphi)$ for $\chi \in H^{-1/2}(\Gamma)$.

For the BIE of the second kind (59), we have the following weak formulation:

Definition 2. Given $\varphi \in H^{1/2}(\Gamma)$, find a function $\mu \in H^{1/2}(\Gamma)$ such that

$$a_{2\Gamma}(\nu, \mu) = \ell_{D\varphi}(\nu) \quad \text{for all } \nu \in H^{1/2}(\Gamma) \quad (69)$$

where the linear functional is defined by $\ell_{D\varphi}(\nu) := (D\nu, \varphi)$ for all $\nu \in H^{1/2}(\Gamma)$.

Also, let us briefly consider the interior Neumann problem,

$$\begin{aligned} \Delta u &= 0 \quad \text{in } \Omega \subset \mathbb{R}^3 \quad \text{and} \quad \frac{\partial u}{\partial n}|_\Gamma = \sigma \quad \text{on } \Gamma \\ \text{with } (\sigma, 1) &= 0 \end{aligned} \quad (70)$$

where $\sigma = (\partial/\partial n)\tilde{u}|_\Gamma$ for some given $\tilde{u} \in H^1(\Omega)$. If we seek the solution

$$u = -W\mu \quad \text{in } \Omega \quad \text{or } \Omega^c \quad (71)$$

in the form of a double-layer potential, then, for its jump across Γ ,

$$\mu = u^- - u^+ = [u]|_\Gamma \quad (72)$$

we arrive at the hypersingular integral equation of the first kind,

$$D\mu = \sigma \quad \text{on } \Gamma \quad (73)$$

Consequently, with u given by (71), we have from (72) the relation

$$\int_\Gamma \mu \overline{D\mu} \, ds = \int_{\Omega \cup \Omega^c} |\nabla u(x)|^2 \, dx$$

If the solution is sought in the form of a simple-layer potential

$$u = V\lambda \quad \text{in } \Omega \quad (74)$$

then this leads to the integral equation of the second kind

$$(\tfrac{1}{2}I + K')\lambda = \sigma \quad \text{on } \Gamma \quad (75)$$

Here, with u given by (74), we find the relation

$$\begin{aligned} \int_\Gamma V\lambda(y)(\tfrac{1}{2}I + K')\overline{\lambda(y)} \, ds &= \int_\Gamma \frac{\partial u}{\partial n} \overline{u} \, ds \\ &= \int_\Omega |\nabla u(x)|^2 \, dx \end{aligned} \quad (76)$$

Now the respective weak formulations corresponding to (73) and (75) read as follows:

Definition 3. Given $\sigma \in H_0^{(-1/2)}(\Gamma) := \{\chi \in H^{(-1/2)}(\Gamma) \mid (\chi, 1) = 0\}$, find $\mu \in H^{(1/2)}(\Gamma)$ such that

$$a_{3\Gamma}(\nu, \mu) := (\nu, D\mu) = (\nu, \sigma) \quad \text{for all } \nu \in H^{(1/2)}(\Gamma) \quad (77)$$

Corresponding to (75), the weak form can be formulated as follows:

Definition 4. Find $\lambda \in H^{(-1/2)}(\Gamma)$ such that

$$\begin{aligned} a_{4\Gamma}(\chi, \lambda) &:= (V\chi, (\tfrac{1}{2}I + K')\lambda) = (V\chi, \sigma) \\ \text{for all } \chi &\in H^{(-1/2)}(\Gamma) \end{aligned} \quad (78)$$

Again, both the sesquilinear forms $a_{3\Gamma}$ and $a_{4\Gamma}$ are symmetric (or Hermitian).

In view of these definitions, we see that the boundary energy spaces are, respectively, $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$ for the above four variational formulations. Indeed, this is true and will be justified in the remaining section; moreover, we present the basic results concerning the existence and uniqueness of the weak solutions of the corresponding BIEs. It is not difficult to see that for the treatment of the elasticity problems, one may follow a similar approach as described here for the Laplacian.

3.2 Basic results

In the weak formulations of BIEs for the model problem (56) given by (68) and (69), we have tacitly assumed the continuity properties of the boundary integral operators involved. In particular, we need that the following operators (corresponding to (16)–(19) in elasticity) are continuous:

$$\begin{aligned} V: H^{-1/2}(\Gamma) &\rightarrow H^{1/2}(\Gamma) \\ D: H^{1/2}(\Gamma) &\rightarrow H^{-1/2}(\Gamma) \\ \tfrac{1}{2}I - K: H^{1/2}(\Gamma) &\rightarrow H^{1/2}(\Gamma) \\ \tfrac{1}{2}I + K': H^{-1/2}(\Gamma) &\rightarrow H^{-1/2}(\Gamma) \end{aligned}$$

These properties can be established by the following theorem.

Theorem 2. Let V and W be the simple- and double-layer potentials defined by (61) and (64) respectively. Then, the following operators are continuous:

$$\begin{aligned} V: H^{(-1/2)}(\Gamma) &\rightarrow H^1(\Omega, \Delta) \otimes H_{\text{loc}}^1(\Omega^c, \Delta) \\ \gamma_0 V \text{ and } \gamma_{\partial\Omega} V: H^{(-1/2)}(\Gamma) &\rightarrow H^{(1/2)}(\Gamma) \\ \tau V \text{ and } \tau_{\partial\Omega} V: H^{(-1/2)}(\Gamma) &\rightarrow H^{(-1/2)}(\Gamma) \\ W: H^{(1/2)}(\Gamma) &\rightarrow H^1(\Omega, \Delta) \otimes H_{\text{loc}}^1(\Omega^c, \Delta) \\ \gamma_0 W \text{ and } \gamma_{\partial\Omega} W: H^{(1/2)}(\Gamma) &\rightarrow H^{(1/2)}(\Gamma) \\ \tau W \text{ and } \tau_{\partial\Omega} W: H^{(1/2)}(\Gamma) &\rightarrow H^{(-1/2)}(\Gamma) \end{aligned}$$

Some explanations for the notations are needed; the function spaces are defined as follows:

$$H^1(\Omega, \Delta) := \{u \in H^1(\Omega) \mid \Delta u \in \tilde{H}^{-1}(\Omega)\}$$

equipped with the graph norm

$$\|u\|_{H^1(\Omega, \Delta)}^2 := \|u\|_{H^1(\Omega)}^2 + \|\Delta u\|_{\tilde{H}^{-1}(\Omega)}^2$$

and the local function space

$$H_{\text{loc}}^1(\Omega^c, \Delta) := \{u \in H_{\text{loc}}^1(\Omega^c) \mid \Delta u \in \tilde{H}_{\text{loc}}^{-1}(\Omega^c)\}$$

We recall that $u \in H_{\text{loc}}^1(\Omega^c)$ (and, respectively, $\Delta u \in \tilde{H}_{\text{loc}}^{-1}(\Omega^c)$) if and only if $\phi u \in H^1(\Omega^c)$ (and $\phi \Delta u \in \tilde{H}^{-1}(\Omega^c)$) for every $\phi \in C_0^\infty(\mathbb{R}^d)$, where $C_0^\infty(\mathbb{R}^d)$ is the space of all infinitely differentiable functions that have compact support in \mathbb{R}^d . The operators γ_0 and $\gamma_{\partial\Omega}$ are the so-called trace operators and have been introduced previously; $\gamma_0 u = u|_\Gamma =: u^-$ for $u \in C^0(\bar{\Omega})$, while $\gamma_{\partial\Omega} u = u|_\Gamma =: u^+$ for $u \in C^0(\bar{\Omega}^c)$. For a smooth function, τu and $\tau_{\partial\Omega} u$ coincide with the normal derivatives $(\partial/\partial n)u^-$ and $(\partial/\partial n)u^+$

respectively. These are linear mappings whose extensions will be given more precisely from the generalized first Green formula later. In terms of the trace operator γ_0 and the linear mapping τ , we have the relations

$$\begin{aligned} \gamma_0 V &= V, & -\tau W &= D \\ -\gamma_0 W &= \tfrac{1}{2}I - K, & \tau V &= \tfrac{1}{2}I + K' \end{aligned}$$

and Theorem 2 provides the continuity of these boundary integral operators. The proof of Theorem 2 will be deferred to the end of this section after we have introduced more machinery.

As a consequence of the mapping properties, we have the following jump relations.

Lemma 2. Given $(\sigma, \varphi) \in H^{(-1/2)}(\Gamma) \times H^{(1/2)}(\Gamma)$, then the following jump relations hold:

$$\begin{aligned} [\gamma_0 V \sigma]_\Gamma &= 0, & [\tau V \sigma]_\Gamma &= \sigma \\ [\gamma_0 W \varphi]_\Gamma &= \varphi, & [\tau W \varphi]_\Gamma &= 0 \end{aligned}$$

Again, the proof will be given later. With respect to the existence of solutions of the variational equations (68) and (69), we have the following results:

Theorem 3. (a) The bilinear form $a_{1\Gamma}$ defined by (68) is $H^{-1/2}(\Gamma)$ -elliptic, that is, there exists a constant $\alpha_1 > 0$ such that the inequality

$$a_{1\Gamma}(\chi, \chi) \geq \alpha_1 \|\chi\|_{H^{-1/2}(\Gamma)}^2 \quad (79)$$

holds for all $\chi \in H^{-1/2}(\Gamma)$. (For $n = 2$, the two-dimensional problems, one needs for (79) appropriate scaling of \mathbb{R}^2 .)

(b) The bilinear forms $a_{j\Gamma}$, $j = 2, 3, 4$, defined by (69), (77), and (78) satisfy Gårding inequalities of the form

$$\operatorname{Re} \{a_{j\Gamma}(\nu, \nu) + (C_j \nu, \nu)\} \geq \alpha_j \|\nu\|_{H^{1/2}(\Gamma)}^2$$

$$\text{for } j = 2, 3 \quad \text{and all } \nu \in H^{1/2}(\Gamma)$$

$$\operatorname{Re} \{a_{4\Gamma}(\lambda, \lambda) + (C_4 \lambda, \lambda)\} \geq \alpha_4 \|\lambda\|_{H^{(-1/2)}(\Gamma)}^2$$

$$\text{for all } \lambda \in H^{(-1/2)}(\Gamma)$$

where

$$C_2, C_3: H^{1/2}(\Gamma) \rightarrow H^{-1/2}(\Gamma)$$

and

$$C_4: H^{-1/2}(\Gamma) \rightarrow H^{1/2}(\Gamma)$$

are compact linear operators, $\alpha_j > 0$ are constants and (\cdot, \cdot) denotes the duality pairing between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$.

In addition, the bilinear forms a_{2c} and a_{3c} are $H_0^{1/2}(\Gamma)$ -elliptic, that is,

$$a_{jc}(\mu, \mu) \geq \alpha_j \|\mu\|_{H^{1/2}(\Gamma)}^2 \quad (80)$$

for all $\mu \in H_0^{j/2}(\Gamma) := \{v \in H^{1/2}(\Gamma) \mid (v, 1) = 0\}$ and $j = 2, 3$; and a_{1c} is $H_0^{-1/2}(\Gamma)$ elliptic, that is,

$$a_{1c}(\lambda, \lambda) \geq \alpha_4 \|\lambda\|_{H^{-1/2}(\Gamma)}^2 \quad \text{for all } \lambda \in H_0^{-1/2}(\Gamma)$$

Compact operators play an important role in our analysis. We include here the basic definition of a linear compact operator.

Definition 5. A linear operator C between two Hilbert spaces X and Y is compact (or completely continuous) iff the image CB of any bounded set $B \subset X$ is a relatively compact subset of Y (i.e. the closure \overline{CB} is compact in Y).

The well-known Lax–Milgram theorem provides a constructive existence proof for any bilinear form, which is elliptic, as a_{1c} . In view of Theorem 3(a), we conclude that the variational equation (68) has a unique solution $\lambda \in H^{-1/2}(\Gamma)$. On the other hand, for bilinear forms such as a_{jc} , $j = 2, 3, 4$, which satisfy Gårding inequalities, the Fredholm theorem then gives the necessary and sufficient conditions for the solvability of the corresponding variational equations. It is not difficult to see that the homogeneous equations of (69) and (77) have nontrivial solutions, $\mu = \text{constants}$, although the corresponding homogeneous equation (59) has only the trivial solution. As in Section 2.3, we may augment (69) and (77) by adding a normalization condition and consider a modified system such as

$$a_{jc}(\nu, \mu) + \omega(\nu, 1) = \ell(\nu) \quad \text{for all } \nu \in H^{1/2}(\Gamma) \quad (81)$$

and

$$(1, \mu) = 0, \quad j = 2 \text{ or } 3$$

where $\omega \in \mathbb{R}$ is now viewed as an unknown constant. In FEM, this is a familiar, so-called mixed formulation; see Hsiao (2000) and Steinbach (2002).

Other equivalent formulations with stabilized symmetric coercive variational sesquilinear forms

$$a_{jc}(\nu, \mu) + (\nu, 1)(1, \mu) = \ell(\nu) \quad \text{for all } \nu \in H^{1/2}(\Gamma), \quad \text{where } j = 2 \text{ or } 3 \quad (82)$$

$$a_{1c}(\chi, \lambda) + (\chi, 1)(1, \lambda) = (\chi, V\sigma) \quad \text{for all } \chi \in H^{-1/2}(\Gamma) \quad (83)$$

(see Fischer *et al.* (1985) and Kuhn and Steinbach (2002)). We note that the variational equation of (69) and those

of (81) and (82) are equivalent. We now summarize our results.

Theorem 4. (a) Equations (68) and (83) as well as equation (82) for $j = 2$ and 3 have unique solutions $\lambda \in H^{-1/2}(\Gamma)$, $\mu \in H^{1/2}(\Gamma)$, respectively. (b) The systems (81) for $j = 2, 3$ have unique solution pairs $(\mu, \omega) \in H^{1/2}(\Gamma) \times \mathbb{R}$. Consequently, (59) admits a unique solution of the form

$$\mu = \mu_0 - c$$

where μ_0 is the unique solution of (81) or (82) for $j = 2$, respectively, and c is the constant defined by

$$c = -\frac{1}{m(\Gamma)} \int_{\Gamma} (\varphi + K\mu_0) \, ds \quad \text{with } m(\Gamma) = \int_{\Gamma} ds$$

We remark that the constant in the theorem plays the same role as an integration constant in the Fichera method (see e.g. Hsiao and MacCamy, 1973), since one may rewrite the variational equation of (81) in the form

$$(D(\tfrac{1}{2}I - K)\mu, v) + \omega(1, v) = (D\varphi, v) \quad \text{for all } v \in H^{1/2}(\Gamma) \quad (84)$$

and both φ and $\varphi + c$ lead to the same right-hand side of (69), (81), and (82). Similarly, for the Neumann problem, the augmented variational system corresponding to (77),

$$a_{2c}(\nu, \mu) + \omega(\nu, 1) = (\nu, \sigma) \quad \text{for all } \nu \in H^{1/2}(\Gamma) \quad (85)$$

together with $(1, \mu) = 0$

or the one corresponding to (78),

$$a_{1c}(\chi, \lambda) + \omega(\chi, 1) = (\chi, V\sigma) \quad \text{for all } \chi \in H^{(-1/2)}(\Gamma) \quad (86)$$

with $(1, \lambda) = 0$

have unique solution pairs $(\mu, \omega) \in H^{1/2}(\Gamma) \times \mathbb{R}$ and $(\lambda, \omega) \in H^{(-1/2)}(\Gamma) \times \mathbb{R}$, correspondingly; and the associated stabilized symmetric variational equations have unique solutions as well. These will be particular solutions of (73) or (75) respectively.

The proof of Theorem 3 will be given after we collect some relevant mathematical tools. The proof of Theorem 4 is a consequence of Theorem 3 together with Theorem 6 and Theorem 7 below.

Since V is $H^{(-1/2)}(\Gamma)$ -elliptic, we may introduce the corresponding energy norm on this space defined by

$$\|\lambda\|_{EV} := (V\lambda, \lambda)^{1/2}$$

which is equivalent to the $H^{(-1/2)}(\Gamma)$ -norm, and its L_2 -dual norm defined by

$$\|\mu\|_{ED} := \sup_{0 \neq \lambda \in H^{(-1/2)}(\Gamma)} \left\{ \frac{|\langle \lambda, \mu \rangle|}{\|\lambda\|_{EV}} \right\}$$

Then one finds that $\|\mu\|_{ED} = (V^{-1}\mu, \mu)^{1/2}$ is equivalent to the $H^{1/2}(\Gamma)$ -norm of μ . The following theorem was shown by Steinbach and Wendland (2001) for Lipschitz boundary Γ and also for the equations of elasticity in Section 2.

Theorem 5. The operators $(1/2)I + K$ and $(1/2)I + K'$ are contractions in the energy spaces and

$$\left\| \left(\tfrac{1}{2}I \pm K \right) \mu \right\|_{ED} \leq c_K \|\mu\|_{ED}$$

$$\text{for all } \mu \in \begin{cases} H^{1/2}(\Gamma) \\ H_0^{1/2}(\Gamma) \end{cases}$$

$$\left\| \left(\tfrac{1}{2}I \pm K' \right) \lambda \right\|_{EV} \leq c_K \|\lambda\|_{EV}$$

$$\text{for all } \lambda \in \begin{cases} H^{(-1/2)}(\Gamma) \\ H_0^{(-1/2)}(\Gamma) \end{cases}$$

where $c_K = (1/2)(1 + \sqrt{1 - 4\alpha_1\alpha_2}) < 1$ and where the respective upper cases correspond to the $+$ signs and the lower ones to the $-$ signs.

This theorem implies that the classical integral equations of the second kind, (33) and (37), can be solved by employing Carl Neumann's classical iterations:

$$\sigma^{(k+1)} := \left(\tfrac{1}{2}I + K' \right) \sigma^{(k)} + D\varphi \quad \text{in } H^{(-1/2)}(\Gamma)$$

and

$$\varphi^{(k+1)} := \left(\tfrac{1}{2}I - K \right) \varphi^{(k)} + V\sigma \quad \text{in } H_0^{1/2}(\Gamma)$$

which converge with respect to the corresponding energy norms $\|\cdot\|_{ED}$ and $\|\cdot\|_{EV}$, respectively.

3.3 Mathematical ingredients

In order to establish the theorems stated above, we need some mathematical tools. In this subsection, we intend to collect the most basic results that we need. For ease of reading, we try to keep the notation simple and the presentation straightforward, although some of the material included may not be in the most general form.

First, we observe that the variational formulations of BVPs for PDEs as well as for BIEs all lead to an abstract variational problem in a Hilbert space \mathcal{H} of the form

Find an element $u \in \mathcal{H}$ such that

$$a(v, u) = \ell(v) \quad \text{for all } v \in \mathcal{H} \quad (87)$$

Here, $a(v, u)$ is a continuous sesquilinear form on \mathcal{H} and $\ell(v)$ is a given continuous linear functional on \mathcal{H} . In order to obtain existence results, one needs to show that $a(\cdot, \cdot)$ satisfies a Gårding inequality in the form

$$\operatorname{Re}[a(v, v) + (Cv, v)_{\mathcal{H}}] \geq \alpha_0 \|v\|_{\mathcal{H}}^2 \quad (88)$$

for all $v \in \mathcal{H}$, where $\alpha_0 > 0$ is a constant and $C: \mathcal{H} \rightarrow \mathcal{H}$ is a compact linear operator. In the most ideal case, when the compact operator vanishes, $C = 0$, the sesquilinear form $a(\cdot, \cdot)$ is said to be \mathcal{H} -elliptic. In this case, we then have the celebrated Lax–Milgram lemma available for the existence proof of the solution, although in most of the cases $C \neq 0$. However, Gårding's inequality implies the validity of the Fredholm alternative, which means that uniqueness implies existence. Since these results are so fundamental, we will state them for the convenience of the reader. We begin with the definition of a sesquilinear form.

Definition 6. A map $a(\cdot, \cdot): \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{C}$ is called a sesquilinear form if it is linear in the first variable and antilinear (i.e. conjugate-linear) in the second. The sesquilinear form $a(\cdot, \cdot)$ is said to be continuous if it satisfies the inequality

$$|a(u, v)| \leq M \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}} \quad \text{for all } u, v \in \mathcal{H}$$

We now state the Lax–Milgram Lemma (see Lax and Milgram, 1954), which, for symmetric $a(\cdot, \cdot)$, is a slight generalization of the Riesz representation theorem using sesquilinear forms instead of the scalar product.

Theorem 6 (The Lax–Milgram Lemma) Let $a(\cdot, \cdot)$ be a continuous \mathcal{H} -elliptic sesquilinear form. Then, to every bounded linear functional $\ell(\cdot)$ on \mathcal{H} , there exists a unique solution $u \in \mathcal{H}$ of the variational equation (87).

In the context of variational problems, if the sesquilinear form $a(\cdot, \cdot)$ satisfies the Gårding inequality (88), then from the Riesz–Schauder Theorem, one may establish the Fredholm alternative for the variational equation (87), which generalizes Theorem 6 (the Lax–Milgram Lemma). In order to state the result, we need the formulations of the homogeneous variational problem and its corresponding adjoint problem for the sesquilinear form $a(\cdot, \cdot)$ on \mathcal{H} . The homogeneous variational problem is to find a solution $u_0 \in \mathcal{H}$ such that

$$a(v, u_0) = 0 \quad \text{for all } v \in \mathcal{H} \quad (89)$$

whereas the corresponding adjoint homogeneous variational problem is to find a solution $v_0 \in \mathcal{H}$ satisfying

$$a^*(u, v_0) := \bar{a}(v_0, u) = 0 \quad \text{for all } u \in \mathcal{H} \quad (90)$$

We also need the adjoint nonhomogeneous problem to (87): Find $v \in \mathcal{H}$ such that

$$a^*(u, v) = \bar{a}(v, u) = \ell^*(u) \quad \text{for all } u \in \mathcal{H} \quad (91)$$

Theorem 7 (Fredholm's alternative) For the variational problem (87), if the sesquilinear form $a(\cdot, \cdot)$ is continuous and satisfies the Gårding inequality (88), then there holds the alternative: either (87) has exactly one solution $u \in \mathcal{H}$ for every given continuous linear functional ℓ on \mathcal{H} or the homogeneous variational problems, (89) and (90), have finite-dimensional eigenspaces of the same dimension. In the latter case, the nonhomogeneous variational problem (87) and its adjoint problem (91) have solutions iff the following orthogonal conditions

$$\ell(v_0) = 0, \quad \text{and} \quad \ell^*(u_0) = 0$$

hold for all the eigensolutions, v_0 of (90) and u_0 of (89) respectively.

For a proof of Theorem 7, see Bers et al. (1964) and Hildebrandt and Wienholtz (1964).

Lemma 3. Let the sesquilinear form $a(\cdot, \cdot)$ satisfy the Gårding inequality (88) on \mathcal{H} , and, in addition,

$$\operatorname{Re} a(v, v) > 0 \quad \text{for all } v \in \mathcal{H} \setminus \{0\}$$

Then, $a(\cdot, \cdot)$ is \mathcal{H} -elliptic.

A contradictory proof of the lemma follows from standard functional analytic arguments.

From our model problems (56) and (70), we see that there is an intimate relation between the sesquilinear forms of the BIEs and that of the underlying PDEs through Green's formula. Indeed, in particular, we see that the linear mappings τ and τ_c employed in Theorem 2 will be defined from the following so-called generalized first Green formulas.

Lemma 4 (Generalized First Green's Formula) For fixed $u \in H^1(\Omega, \Delta)$, the mapping

$$v \mapsto (\tau u, v)_\Gamma := a_\Omega(u, Zv) + \int_\Omega (\Delta u) \overline{Zv} \, dx$$

is a continuous antilinear functional τu on $v \in H^{1/2}(\Gamma)$ that coincides for $u \in H^2(\Omega)$ with $(\partial u / \partial n)$, that is, $\tau u =$

$(\partial u / \partial n)$. The linear mapping $\tau : H^1(\Omega, \Delta) \rightarrow H^{(-1/2)}(\Gamma)$ with $u \mapsto \tau u$ is continuous. Here, Z is a right inverse to the trace operator γ_0 . Thus, there holds the generalized first Green's formula

$$-(\Delta u, v)_\Omega = - \int_\Omega (\Delta u) \overline{v} \, dx = a_\Omega(u, v) - (\tau u, \gamma_0 v)_\Gamma \quad (92)$$

for $u \in H^1(\Omega, \Delta)$ and $v \in H^1(\Omega)$. Here, $a_\Omega(u, v) := \int_\Omega \nabla u \cdot \overline{\nabla v} \, dx$ is a sesquilinear form and (\cdot, \cdot) denotes the duality pairings such that

$$(\tau u, \gamma_0 v)_\Gamma = (\tau u, \gamma_0 v)_{L^2(\Gamma)} = \int_\Gamma (\tau u) \overline{\gamma_0 v} \, dx$$

provided $\tau u \in H^{(-1/2)}(\Gamma)$ and $\gamma_0 v \in H^{1/2}(\Gamma)$.

Lemma 5 (Exterior Generalized First Green's Formula) Let $u \in H_{\text{loc}}^1(\Omega^c, \Delta)$ be given. Then, the linear functional $\tau_c u$ defined by the mapping

$$(\tau_c u, \gamma_{\text{co}} \lambda)_\Gamma := -a_{\Omega^c}(u, Z_c \lambda) - \int_{\Omega^c} \Delta u \overline{Z_c \lambda} \, dx$$

$$\text{for all } \lambda \in H^{1/2}(\Gamma)$$

belongs to $H^{(-1/2)}(\Gamma)$ and coincides with $(\partial u / \partial n)$, provided $u \in C^2(\overline{\Omega^c})$. Moreover, $\tau_c : u \mapsto \tau_c u$ is a linear continuous mapping from $H_{\text{loc}}^1(\Omega^c, \Delta)$ into $H^{(-1/2)}(\Gamma)$, and there holds the exterior generalized first Green's formula

$$- \int_{\Omega^c} (\Delta u) \overline{v} \, dx = a_{\Omega^c}(u, v) + (\tau_c u, \gamma_{\text{co}} v)_\Gamma \quad (93)$$

for $u \in H_{\text{loc}}^1(\Omega^c, \Delta)$ and for every $v \in H_{\text{comp}}^1(\overline{\Omega^c})$.

In Lemma 5, the function space $H_{\text{comp}}^1(\overline{\Omega^c})$ is defined by

$$H_{\text{comp}}^1(\overline{\Omega^c}) := \{v \in H_{\text{loc}}^1(\Omega^c) \mid v \text{ has compact support in } \mathbb{R}^3\}$$

The operator Z_c denotes the right inverse to γ_{co} , which maps $v \in H^{1/2}(\Gamma)$ into $Z_c v \in H_{\text{comp}}^1(\overline{\Omega^c})$ with the $\operatorname{supp}(Z_c v)$ contained in some fixed compact set $\Omega_R^c \subset \mathbb{R}^3$ containing Γ . The sesquilinear form $a_{\Omega^c}(u, v)$ and the duality pairing $(\cdot, \cdot)_\Gamma$ are defined similarly as those in Lemma 4.

Generalized first Green's formulas give the relations between the solution spaces of weak solutions of the PDEs and the corresponding boundary integral equations. We also need the generalized formula for deriving the representation formula for the weak solutions. This is based on the generalized second Green formula.

Lemma 6 (Generalized Second Green's Formula) For every pair $u, v \in H^1(\Omega, \Delta)$, there holds the formula

$$\int_\Omega (u(\Delta v) - \Delta u \overline{v}) \, dx = ((\tau u), \gamma_0 v)_\Gamma - ((\gamma_0 u), \tau v)_\Gamma$$

For the proofs of generalized Green formulas and for the systematic study of Sobolev spaces, the reader is referred to Nečas (1967) or Lions and Magenes (1972), and also Aubin (1972).

3.4 Proofs of basic results

For the proofs of Theorem 2 and Lemma 2, we need a representation formula for the variational solution of the transmission problem for

$$-\Delta u = 0 \quad \text{in } \mathbb{R}^3 \setminus \Gamma \quad (94)$$

satisfying the decay condition $u = O(|x|^{-1})$ as $|x| \rightarrow \infty$.

Theorem 8 (Generalized Representation Formula) Let u be a variational solution for the transmission problem (94) with $u|_\Omega \in H^1(\Omega, \Delta)$ and $u|_{\Omega^c} \in H_{\text{loc}}^1(\Omega^c, \Delta)$. Then, $u(x)$ admits the representation

$$u(x) = ((\tau u)_\Gamma, E(x, \cdot))_\Gamma - \left\langle \frac{\partial}{\partial n} E(x, \cdot), [\gamma_0 u]_\Gamma \right\rangle_\Gamma$$

$$\text{for } x \in \mathbb{R}^3 \setminus \Gamma \quad (95)$$

where $[\gamma_0 u]_\Gamma$ and $(\tau u)_\Gamma$ denote the jumps of $\gamma_0 u$ and τu across Γ respectively.

For problems in the plane, $n = 2$, this representation formula is to be modified slightly because of the logarithmic behavior of the simple-layer potential.

Proof. For the proof, we shall use the following estimate:

$$|((\tau u)_\Gamma, E(x, \cdot))_\Gamma| + \left| \left\langle \frac{\partial}{\partial n} E(x, \cdot), [\gamma_0 u]_\Gamma \right\rangle_\Gamma \right| \leq c_{\varepsilon, R} \left\{ \|u\|_{H^1(\Omega, \Delta)} + \|u\|_{H^1(\Omega_R^c, \Delta)} \right\} \quad (96)$$

where $R > 0$ is sufficiently large so that $\Omega_R^c = \Omega^c \cap \{|y| \in \mathbb{R}^3 \mid |y| < R\}$ contains $\overline{\Omega}$. Since $E(x, \cdot)$ for fixed $x \notin \Gamma$ is a smooth function on the boundary Γ , the estimate (96) follows from Lemmas 4 and 5 together with the continuity of the duality $(\cdot, \cdot)_\Gamma$.

With the estimate (96), the representation (95) can be established by the usual completing procedure. More precisely, if u is a variational solution with the required properties, we can approximate u by a sequence of functions u_k

with

$$u_k|_\Omega \in C^\infty(\overline{\Omega}) \quad \text{and} \quad u_k|_{\Omega^c} \in C_{\text{comp}}^\infty(\overline{\Omega^c})$$

satisfying the classical representation formula,

$$u_k(x) = \int_\Gamma E(x, y) \left[\frac{\partial u_k}{\partial n} \right]_\Gamma \, ds_y - \int_\Gamma \frac{\partial}{\partial n_y} E(x, y) [u_k]_\Gamma \, ds_y$$

$$\text{for } x \in \mathbb{R}^3 \setminus \Gamma \quad (97)$$

so that

$$\|u - u_k\|_{H^1(\Omega, \Delta)} + \|u - u_k\|_{H_{\text{loc}}^1(\Omega^c, \Delta)} \rightarrow 0 \quad \text{as } k \rightarrow \infty$$

Then, because of (96), for any $x \notin \Gamma$, the two boundary potentials in (97) generated by u_k will converge to the corresponding boundary potentials in (95). This completes the proof. \square

Observe that we may rewrite the representation (95) in the form

$$u(x) = V[(\tau u)_\Gamma](x) - W[\gamma_0 u]_\Gamma(x) \quad \text{for } x \in \mathbb{R}^3 \setminus \Gamma \quad (98)$$

where $V\sigma(x)$ is the simple-layer potential and $W\varphi(x)$ is the double-layer potential, namely,

$$V\sigma(x) := (E(x, \cdot), \sigma)_\Gamma$$

$$W\varphi(x) := \left\langle \frac{\partial}{\partial n} E(x, \cdot), \varphi \right\rangle_\Gamma \quad (99)$$

From this formula (98), we may now establish the mapping properties in Theorem 2.

Proof of Theorem 2 For the continuity of V , we consider the transmission problem

Find $u \in H^1(\Omega, \Delta)$ and $u \in H_{\text{loc}}^1(\Omega^c, \Delta)$ satisfying the differential equation

$$-\Delta u = 0 \quad \text{in } \Omega \text{ and } \Omega^c$$

together with the transmission conditions

$$[\gamma_0 u]_\Gamma = 0 \quad \text{and} \quad (\tau u)_\Gamma = \sigma$$

with given $\sigma \in H^{(-1/2)}(\Gamma)$ and the decay condition $u = O(|x|^{-1})$ as $|x| \rightarrow \infty$.

This variational transmission problem has a unique solution u . Moreover, u depends on σ continuously. The latter implies that the mappings $\sigma \mapsto u$ from $H^{(-1/2)}(\Gamma)$ to $H^1(\Omega, \Delta)$ and to $H_{\text{loc}}^1(\Omega^c, \Delta)$, respectively, are continuous. By the representation (99), it follows that $\sigma \mapsto u = V\sigma$ is

continuous in the corresponding spaces. This, together with the continuity of the trace operators

$$\gamma_0: H^1(\Omega^c, \Delta) \rightarrow H^{1/2}(\Gamma)$$

and

$$\gamma_{0*}: H_{loc}^1(\Omega^c, \Delta) \rightarrow H^{1/2}(\Gamma)$$

and of the linear mappings (from Lemmas 4 and 5)

$$\tau: H^1(\Omega^c, \Delta) \rightarrow H^{-1/2}(\Gamma)$$

and

$$\tau_c: H_{loc}^1(\Omega^c, \Delta) \rightarrow H^{-1/2}(\Gamma)$$

implies the continuity properties of the operators associated with V .

Similarly, the solution of the transmission problem Find $u \in H^1(\Omega, \Delta)$ and $u \in H_{loc}^1(\Omega^c, \Delta)$ satisfying

$$-\Delta u = 0 \text{ in } \Omega \text{ and in } \Omega^c$$

with the transmission conditions

$$[\gamma_0 u]_\Gamma = \varphi \in H^{1/2}(\Gamma) \text{ and } [\tau u]_\Gamma = 0$$

together with the representation (99),

$$u(x) = -W\varphi(x) \text{ for } x \in \mathbb{R}^3 \setminus \Gamma$$

provides the desired continuity properties for the mappings associated with W .

Proof of Lemma 2 We see from the representation (98) that $u(x) = V\sigma(x)$ for $x \in \mathbb{R}^3 \setminus \Gamma$ is the solution of the transmission problem for (94) with the transmission conditions

$$[\gamma_0 u]_\Gamma = 0 \text{ and } \sigma = [\tau u]_\Gamma$$

Inserting $u = V\sigma$ gives the jump relations involving V . Likewise, $u(x) = -W\varphi(x)$ is the solution of the transmission problem of (94) satisfying

$$[\gamma_0 u]_\Gamma = \varphi \text{ and } [\tau u]_\Gamma = 0$$

which gives the desired jump relations involving W . This completes the proof of Lemma 2.

Remark 1. In accordance with the classical formulations in the Hölder spaces, we now introduce the boundary integral operators on Γ for given $(\sigma, \varphi) \in H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)$ defined by

$$\begin{aligned} V\sigma &:= \gamma_0 V\sigma, & K\varphi &:= \gamma_0 W\varphi + \frac{1}{2}\varphi \\ K'\sigma &:= \tau V\sigma - \frac{1}{2}\sigma, & D\varphi &:= -\tau W\varphi \end{aligned}$$

Clearly, Lemma 4 provides us with the continuity of these boundary integral operators. The corresponding Calderon projectors can now be defined in a weak sense by

$$C_\Omega := \begin{pmatrix} \frac{1}{2}I - K & V \\ D & \frac{1}{2}I + K' \end{pmatrix}$$

and

$$C_{\Omega^c} := I - C_\Omega$$

which are continuous mappings on $(H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma))$.

To establish Theorem 3, we need Lemmas 4, 5, and 2 together with the Gårding inequality for the sesquilinear form for the underlying partial differential operator. For the model problem with homogeneous Dirichlet conditions, we have

$$a_\Omega(u, v) := \int_\Omega \nabla \bar{u} \nabla v \, dx \text{ for all } u, v \in H^1(\Omega) \quad (100)$$

and there holds a Gårding inequality in the form

$$\operatorname{Re} a_\Omega(v, v) \geq \|v\|_{H^1(\Omega)}^2 - \|v\|_{L^2(\Omega)}^2 \quad (101)$$

for all $v \in H^1(\Omega)$. The latter implies that there is a compact linear operator C defined by

$$(Cu, v)_{H^1(\Omega)} := \int_\Omega u \bar{v} \, dx$$

so that (101) can be rewritten as

$$\operatorname{Re} a_\Omega(v, v) + (Cv, v)_{H^1(\Omega)} \geq \alpha_0 \|v\|_{H^1(\Omega)}^2 \quad (102)$$

with some constant $\alpha_0 > 0$. The compactness of C can be seen as follows. Since $(Cu, v)_{H^1(\Omega)} = (u, C^*v)_{H^1(\Omega)}$, and from the estimate

$$|(u, C^*v)_{H^1(\Omega)}| = |(u, v)_{L^2(\Omega)}| \leq \|u\|_{H^1(\Omega)} \|v\|_{L^2(\Omega)}$$

it follows that C^* maps $L^2(\Omega)$ into $H^1(\Omega)$ continuously. Then the Sobolev imbedding theorem implies that $C^*: H^1(\Omega) \rightarrow H^1(\Omega)$ is compact and hence C is compact (see Adams, 1975).

Proof of Theorem 3 The proof follows Hsiao and Wendland (1977) and Costabel and Wendland (1986). For the sesquilinear form $a_{\Gamma}(\cdot, \cdot)$, we first show that it satisfies a Gårding inequality of the form

$$\begin{aligned} \operatorname{Re} a_{\Gamma}(\sigma, \sigma) + (\sigma, C_V \sigma)_{\Gamma} &\geq \alpha_1 \|\sigma\|_{H^{-1/2}(\Gamma)}^2 \\ \text{for all } \sigma &\in H^{-1/2}(\Gamma) \end{aligned} \quad (103)$$

where $C_V: H^{-1/2}(\Gamma) \rightarrow H^{1/2}(\Gamma)$ is compact. Then the $H^{-1/2}(\Gamma) \rightarrow$ ellipticity of $a_{\Gamma}(\cdot, \cdot)$ follows immediately from Lemma 3, since

$$\operatorname{Re} a_{\Gamma}(\sigma, \sigma) := \operatorname{Re} (\sigma, V\sigma) \geq 0 \text{ for all } \sigma \in H^{-1/2}(\Gamma)$$

and

$$\operatorname{Re} (\sigma, V\sigma) = 0 \text{ implies } \sigma = 0$$

as will be shown below. For any $\sigma \in H^{-1/2}(\Gamma)$, let

$$v(x) := \int_{\Gamma} E(x, y) \sigma(y) \, dy \text{ for } x \in \mathbb{R}^3 \setminus \Gamma$$

Then, $v \in H^1(\Omega, \Delta)$ and $v \in H_{loc}^1(\Omega^c, \Delta)$ respectively. Moreover, Lemma 2 yields

$$[\gamma_0 v]_\Gamma = 0 \text{ and } [\tau v]_\Gamma = \sigma$$

By adding the generalized Green formulas (92) and (93), we obtain

$$\begin{aligned} (\sigma, V\sigma)_{\Gamma} &= \int_{\Omega} |\nabla v|^2 \, dx + \int_{\Omega^c} |\nabla v|^2 \, dx \\ &= a_\Omega(v, v) + a_{\Omega^c}(v, v) \end{aligned} \quad (104)$$

The right-hand side of (104) contains two sesquilinear forms for the corresponding partial differential operator $-\Delta$, one in Ω and the other one in the exterior domain Ω^c . It is this relation, in some sense, that connects intimately the Gårding inequality of V with that of $-\Delta$. We see from (100) that

$$a_\Omega(v, v) = \|v\|_{H^1(\Omega)}^2 - (Cv, v)_{H^1(\Omega)}$$

where C is a compact operator from $H^1(\Omega)$ into itself. This takes care of the contribution to the Gårding inequality for V from the interior domain Ω . We would expect, of course, to have a similar result for the exterior domain Ω^c . However, there is a technical difficulty, since v defined by the simple-layer potential is not in $L^2(\Omega^c)$. In order to apply similar arguments to Ω^c , following Costabel and Wendland (1986), we introduce a fixed $C_0^\infty(\mathbb{R}^3)$ cut-off function ϕ with $\phi|_{\Omega} = 1$ and $\operatorname{dist}(\{x \in \mathbb{R}^3 | \phi(x) \neq 1\}, \Omega) =: d_0 > 0$, since the exterior generalized first Green formula in Lemma 6 is valid for $v_c := \phi v$ having compact support. With this modification, (104) becomes

$$\begin{aligned} (\sigma, V\sigma)_{\Gamma} + \int_{\Omega^c} (-\Delta v_c) \bar{v}_c \, dx \\ = \int_{\Omega} |\nabla v|^2 \, dx + \int_{\Omega^c} |\nabla v_c|^2 \, dx =: a_\Omega(v, v) + a_{\Omega^c}(v_c, v_c) \end{aligned} \quad (105)$$

On the other hand, from Lemmas 4 and 6, the mappings $\tau: H^1(\Omega, \Delta) \rightarrow H^{-1/2}(\Gamma)$ and $\tau_c: H_{loc}^1(\Omega^c, \Delta) \rightarrow H^{-1/2}(\Gamma)$ are continuous. Thus, we have

$$\begin{aligned} \|\sigma\|_{H^{-1/2}(\Gamma)}^2 \\ = \|[\tau v]_\Gamma\|_{H^{-1/2}(\Gamma)}^2 \leq 2\|\tau v\|_{H^{-1/2}(\Gamma)}^2 + 2\|\tau_c v\|_{H^{-1/2}(\Gamma)}^2 \\ \leq c \left\{ \|v\|_{H^1(\Omega)}^2 + \|v_c\|_{H^1(\Omega^c)}^2 + \|\Delta v_c\|_{H^{-1}(\Omega^c)}^2 \right\} \end{aligned}$$

where $\Omega' := \Omega^c \cap \operatorname{supp} \phi$. Then the Gårding's inequality (102) for a_Ω in the domain Ω and for a_{Ω^c} in Ω' implies

$$\begin{aligned} \alpha_0 \left\{ \|v\|_{H^1(\Omega)}^2 + \|v_c\|_{H^1(\Omega^c)}^2 \right\} \leq \operatorname{Re} \{ a_\Omega(v, v) + a_{\Omega^c}(v_c, v_c) \\ + (Cv, v)_{H^1(\Omega)} + (C'v_c, v_c)_{H^1(\Omega')} \} \end{aligned}$$

with a positive constant α_0 and compact operators C and C' depending also on Ω and Ω' . Collecting the inequalities, we get

$$\begin{aligned} \alpha_1 \|\sigma\|_{H^{-1/2}(\Gamma)}^2 \leq \operatorname{Re} \{ a_\Omega(v, v) + a_{\Omega^c}(v_c, v_c) \\ + c_1 (Cv, v)_{H^1(\Omega)} + c_2 (C'v_c, v_c)_{H^1(\Omega')} + c_3 \|\Delta v_c\|_{H^{-1}(\Omega^c)}^2 \} \end{aligned} \quad (106)$$

Then, from (105) and (106), we finally obtain the inequality

$$\alpha_1 \|\sigma\|_{H^{-1/2}(\Gamma)}^2 \leq \operatorname{Re} \{ a_{\Gamma}(\sigma, \sigma) + (\sigma, C_V \sigma)_{\Gamma} \}$$

where the operator $C_V: H^{-1/2}(\Gamma) \rightarrow H^{1/2}(\Gamma)$ is defined by the sesquilinear form

$$\begin{aligned} (\chi, C_V \sigma)_{\Gamma} = c(\chi, \sigma) := c_1 (V\chi, C_V \sigma)_{H^1(\Omega)} \\ + c_2 (\phi V\sigma, C' \phi V\sigma)_{H^1(\Omega')} + c_3 (\Delta(\phi V\sigma), \Delta(\phi V\sigma))_{H^{-1}(\Omega^c)} \\ + (-\Delta(\phi V\chi), \phi V\sigma)_{L^2(\Omega^c)} \end{aligned}$$

We note that the operator C_V is well defined, since each term on the right-hand side is a bounded sesquilinear form on $\chi, \sigma \in H^{-1/2}(\Gamma)$. Hence, by the Riesz representation theorem, there exists a linear mapping $jC_V: H^{-1/2}(\Gamma) \rightarrow H^{-1/2}(\Gamma)$ such that

$$c(\chi, \sigma) = (\chi, jC_V \sigma)_{H^{-1/2}(\Gamma)}$$

Since $j^{-1}: H^{-1/2}(\Gamma) \rightarrow (H^{-1/2}(\Gamma))' = H^{1/2}(\Gamma)$ is continuous, this representation can be written in the desired form

$$(\chi, C_V \sigma)_{\Gamma} = c(\chi, \sigma) \text{ for all } \chi, \sigma \in H^{-1/2}(\Gamma)$$

where $C_V = j^{-1} jC_V$. It remains to be shown that C_V is compact. This follows from the standard arguments based

on the Sobolev imbedding theorem and the compactness of the operators C and C' . We refer the reader to Hsiao and Wendland (2005) for details.

To establish the Gårding inequality for the sesquilinear form

$$a_{2\epsilon}(v, \mu) := \langle Dv, (\frac{1}{2}I - K)\mu \rangle$$

we set for fixed $\mu \in H^{1/2}(\Gamma)$,

$$w(x) := - \int_{\Gamma} \frac{\partial}{\partial n_y} E(x, y) \mu(y) \, d\epsilon_y \quad \text{for } x \in \mathbb{R}^3 \setminus \Gamma$$

It follows from Theorem 2, Lemma 2, and the generalized first Green formula in Lemma 4 that we arrive at the integral relation

$$\langle D\mu, (\frac{1}{2}I - K)\mu \rangle_{\Gamma} = a_{\epsilon}(w, \mu) \geq 0 \quad (107)$$

Thus, from (102), we have

$$\begin{aligned} \operatorname{Re} \langle D\mu, (\frac{1}{2}I - K)\mu \rangle_{\Gamma} + \langle Cw, w \rangle_{H^1(\Omega)} &\geq \alpha_0 \|w\|_{H^1(\Omega)}^2 \\ &\geq \alpha_5 \|\tau w\|_{H^{-1/2}(\Gamma)}^2 = \alpha_5 \|D\mu\|_{H^{-1/2}(\Gamma)}^2 \end{aligned} \quad (108)$$

where $\alpha_5 > 0$, by using the continuity of the linear mapping τ in Lemma 4. Now let us assume that we have the Gårding inequality for the operator D in the form

$$\operatorname{Re} \langle \langle \mu, D\mu \rangle_{\Gamma} + \langle \mu, C_D \mu \rangle \rangle \geq \alpha_D \|\mu\|_{H^{1/2}(\Gamma)}^2 \quad (109)$$

where $C_D: H^{1/2}(\Gamma) \rightarrow H^{-1/2}(\Gamma)$ is a compact operator. Then, we have the estimate

$$\|D\mu\|_{H^{-1/2}(\Gamma)}^2 \geq \alpha_D \|\mu\|_{H^{1/2}(\Gamma)}^2 - \|C_D \mu\|_{H^{-1/2}(\Gamma)}^2$$

This together with (108) implies that

$$\operatorname{Re} \langle \langle D\mu, (\frac{1}{2}I - K)\mu \rangle_{\Gamma} + c(\mu, \mu)_{\Gamma} \rangle \geq \alpha_2 \|\mu\|_{H^{1/2}(\Gamma)}^2$$

where

$$c(\mu, \mu)_{\Gamma} := \alpha_5 \langle C_D \mu, C_D \mu \rangle_{H^{-1/2}(\Gamma)} + \langle W\mu, CW\mu \rangle_{H^1(\Omega)}$$

the compactness of which follows from the arguments as before.

To complete the proof, it remains to establish the Gårding inequality (109). The proof is based on the estimate

$$\begin{aligned} \|\mu\|_{H^{1/2}(\Gamma)}^2 &= \|(\gamma_0 w)_{\Gamma}\|_{H^{1/2}(\Gamma)}^2 \\ &\leq c \{ \|w\|_{H^1(\Omega)}^2 + \|\phi w\|_{H^1(\Omega)}^2 \} \end{aligned}$$

which follows from the trace theorem. The remaining arguments of the proof are again the same as in the proof

for the operator V in (105). In fact, (109) is Gårding's inequality for $a_{2\epsilon}$. Because of (76), Gårding's inequality for a_{ϵ} follows in the same way as that for $a_{2\epsilon}$.

For the proof of (80) for $j = 2$, we employ Lemma 3, since $a_{2\epsilon}(\mu_0, \mu_0) = 0$ with (107) implies $w_0 = \text{constant}$ in Ω and, with (6) for w_0 and uniqueness for (59), one finds $\mu_0 = w_0$ on Γ and $\mu_0 = 0$ if $\mu_0 \in H_0^{1/2}(\Gamma)$.

The proof of Theorem 5 rests on the Calderon projection properties and the relations between norms of dualities. For details, we refer the reader to (Steinbach and Wendland, 2001).

Remark 2. We notice that (104) implies the $H^{-1/2}(\Gamma)$ -ellipticity of V immediately if weighted Sobolev spaces were used as in the French school (see Nedelec and Planchard, 1973; LeRoux, 1977). However, we choose not to do so by introducing the cut-off function ϕ . It is also worth mentioning that the Gårding inequality for $a_{2\epsilon}(\cdot, \cdot)$ is established without using compactness of the operator K . In particular, we see that in (108), we may obtain the estimate

$$\|w\|_{H^1(\Omega)}^2 \geq c \|w\|_{H^{1/2}(\Gamma)}^2 \geq c_1 (\|\mu\|_{H^{1/2}(\Gamma)}^2 - \|K\mu\|_{H^{1/2}(\Gamma)}^2)$$

by the trace theorem; if K is compact, the proof will be more direct and shorter without employing (109). However, we deliberately do so without using the above estimate since our approach here can also be applied to the elasticity as well as for a Lipschitz boundary, in which case, also for the Laplacian, K is no longer compact.

4 THE GALERKIN-BEM

This section is devoted to the Galerkin-BEM for the same model problems as in the previous section. From the formulation of the Galerkin system to the basic results of error estimates, it is a section that describes the Galerkin-BEM and provides detailed proofs of basic results for typical BIEs. Needless to say, the approach is general enough and can be adapted to other cases including elasticity with slight or even without any modifications.

4.1 Galerkin equations and Céa's lemma

To simplify the presentation, we begin with the variational equation (68) for the BIE of the first kind (58). The boundary energy space is $H^{-1/2}(\Gamma)$. The Galerkin method consists of seeking an approximate solution of (68) in a finite-dimensional subspace S_h of the space $H^{-1/2}(\Gamma)$ consisting of admissible functions, rather than in the whole space.

More precisely, let S_h be a family of finite-dimensional subspaces that approximate $H^{-1/2}(\Gamma)$, that is,

for every $\lambda \in H^{-1/2}(\Gamma)$, there exists a sequence

$$\lambda_h \in S_h \subset H^{-1/2}(\Gamma)$$

such that $\|\lambda_h - \lambda\|_{H^{-1/2}(\Gamma)} \rightarrow 0$ as $h \rightarrow 0$ (110)

that is, the degrees of freedom $N \rightarrow \infty$. Then the Galerkin approximation of the solution λ of (68) is a function $\lambda_h \in S_h$ satisfying the Galerkin equations

$$a_{1\epsilon}(\chi_h, \lambda_h) = \ell_{\epsilon}(\chi_h) \quad \text{for all } \chi_h \in S_h \subset H^{-1/2}(\Gamma) \quad (111)$$

The Galerkin solution λ_h is in some sense the discrete weak solution of (58). Alternatively, since S_h is finite-dimensional, if $\{\mu_i(x)\}_i^N$ is a basis of S_h , then we may seek a solution λ_h in the form

$$\lambda_h := \sum_{j=1}^N y_j \mu_j(x)$$

where the unknown coefficients are now required to satisfy the linear system

$$\sum_{j=1}^N a_{1\epsilon}(\mu_i, \mu_j) y_j = \ell_{\epsilon}(\mu_i), \quad i = 1, \dots, N \quad (112)$$

As a consequence of the $H^{-1/2}(\Gamma)$ -ellipticity of $a_{1\epsilon}(\cdot, \cdot)$, it is easy to see that the Galerkin equations (111) or (112) are uniquely solvable, and the essential properties concerning the Galerkin approximation λ_h can be stated in the following theorem without specifying the subspace S_h in any particular form.

Theorem 9. *There exists an $h_0 > 0$ such that the corresponding Galerkin equations (111) (or 112) admit a unique solution $\lambda_h \in S_h \subset H^{-1/2}(\Gamma)$ for every $h \leq h_0$. Moreover, the Galerkin projections G_{hV} defined by*

$$G_{hV}: H^{-1/2}(\Gamma) \ni \lambda \mapsto \lambda_h \in S_h \subset H^{-1/2}(\Gamma)$$

are uniformly bounded, that is,

$$\|G_{hV}\| := \sup_{\|\lambda\|_{H^{-1/2}(\Gamma)} \leq 1} \|G_{hV}\lambda\|_{H^{-1/2}(\Gamma)} \leq c \quad (113)$$

for all $h \leq h_0$, where $c = c(h_0)$. Consequently, we have the estimate

$$\begin{aligned} \|\lambda - \lambda_h\|_{H^{-1/2}(\Gamma)} &\leq (1 + c) \inf_{\chi_h \in S_h} \|\lambda - \chi_h\|_{H^{-1/2}(\Gamma)} \rightarrow 0 \\ &\text{as } h \rightarrow 0 \end{aligned} \quad (114)$$

The corresponding result also holds for the Galerkin scheme for a_{ϵ} in (83) and (84).

The estimate (114) is usually known as Céa's Lemma in finite element analysis (see also Ciarlet, 1978). Moreover, (114) provides the basic inequality for obtaining convergence results in norms other than the energy norm for the integral operator V . As in the case of PDEs, this simple yet crucial estimate, (114), shows that the problem of estimating the error between the solution λ and its Galerkin approximation λ_h is reduced to a problem in approximation theory. In particular, if we assume that the finite dimensional subspaces S_h are regular boundary element spaces as introduced in Babuška and Aziz (1977), then one may obtain convergence results with respect to a whole range of Sobolev space norms, including superconvergence results, by using the Aubin–Nitsche lemma for BIEs as in Hsiao and Wendland (1981a,b). We will discuss the details in the next two subsections.

To prove Theorem 9, we first notice that by the $H^{-1/2}(\Gamma)$ -ellipticity of V (79), we have

$$\begin{aligned} |(\lambda_h, V\lambda_h)| &\geq \operatorname{Re}(\lambda_h, V\lambda_h) \geq \alpha_1 \|\lambda_h\|_{H^{-1/2}(\Gamma)}^2 \\ &\quad \text{for all } \lambda_h \in S_h \end{aligned} \quad (115)$$

with some constant $\alpha_1 > 0$, independent of h . This implies that the homogeneous equations of (111) have only the trivial solution in S_h and, hence, (111) has a unique solution. Now we introduce a family of projections $\mathbb{P}_h: H^{(-1/2)}(\Gamma) \rightarrow S_h \subset H^{(-1/2)}(\Gamma)$ that is uniformly bounded, that is, there exists a constant c_p such that

$$\|\mathbb{P}_h\| \leq c_p \quad \text{for all } 0 < h \leq h_0 \quad (116)$$

with some fixed $h_0 > 0$. If \mathbb{P}_h is chosen as the L^2 -orthogonal projection onto S_h , as in Hsiao and Wendland (1977), then (116) is a requirement for the finite element spaces on Γ , as analyzed by Steinbach (2001), which is fulfilled if these provide approximation as well as inverse properties; see also Aubin (1972) and Babuška and Aziz (1977). Now the Galerkin equations (111) are equivalent to the operator equation

$$\mathbb{P}_h^* V \mathbb{P}_h \lambda_h = \mathbb{P}_h^* \varphi = \mathbb{P}_h^* V \lambda \quad (117)$$

where $\mathbb{P}_h^*: H^{(1/2)}(\Gamma) \rightarrow S_h$ denotes the adjoint of \mathbb{P}_h . Various realizations of projections \mathbb{P}_h and \mathbb{P}_h^* are analyzed by Steinbach (2002).

We have shown that this equation has a unique solution, from which one may define the Galerkin projection $G_{hV}: \lambda \mapsto \lambda_h$ such that

$$\lambda_h = G_{hV} \lambda := (\mathbb{P}_h^* V \mathbb{P}_h)^{-1} (\mathbb{P}_h^* V) \lambda \quad (118)$$

Observe that for $\chi_h \in \mathcal{S}_h$, $\chi_h = \mathbb{P}_h \chi_h$, and hence $G_h \chi_h = \chi_h$, that is, $G_h \chi_h = I_{\mathcal{S}_h}$. Moreover, from (115), we see that

$$\|\lambda_h\|_{H^{-1/2}(\Gamma)} \|V\lambda\|_{H^{1/2}(\Gamma)} \geq |(\lambda_h, V\lambda)| = |(\lambda_h, \mathbb{P}_h^* V \mathbb{P}_h \lambda_h)| \\ = |(\lambda_h, V\lambda_h)| \geq \alpha_1 \|\lambda_h\|_{H^{-1/2}(\Gamma)}^2 \quad \text{for all } \lambda_h \in H^{-1/2}(\Gamma)$$

and the continuity of V and (79) imply that the estimates

$$\alpha_1 \|\lambda_h\|_{H^{-1/2}(\Gamma)} = \alpha_1 \|G_h V \lambda\|_{H^{-1/2}(\Gamma)} \leq \|V\lambda\|_{H^{1/2}(\Gamma)} \\ \leq M \|\lambda\|_{H^{-1/2}(\Gamma)}$$

hold for all $\lambda \in H^{-1/2}(\Gamma)$, where M and α_1 are constants independent of h . Then it follows that the Galerkin projection G_h is uniformly bounded, that is, (113) holds with $c := M/\alpha_1$, independent of h and λ for all $h \leq h_0$.

To complete the proof, it remains to establish the estimate (113). This is a consequence of (114) together with the definition (118). This can be seen as follows:

$$\|\lambda - \lambda_h\|_{H^{-1/2}(\Gamma)} = \|(\lambda - \chi_h) + G_h V(\chi_h - \lambda)\|_{H^{-1/2}(\Gamma)} \\ \leq (1+c) \|\chi_h - \lambda\|_{H^{-1/2}(\Gamma)} \quad \text{for every } \chi_h \in \mathcal{S}_h$$

from which (114) results.

Because of the limitation of space, we omit the proof of the corresponding results for a_{2c} .

Now we consider the Galerkin approximation of the solution $(\mu, \omega) \in H^{1/2}(\Gamma) \times \mathbb{R}$ to the augmented system (85) by using a family of finite-dimensional subspaces $\mathcal{B}_h \subset H^{1/2}(\Gamma)$ that approximate $H^{1/2}(\Gamma)$, that is,

$$\text{for every } v \in H^{1/2}(\Gamma), \text{ there exists a sequence } \mu_h \in \mathcal{B}_h \subset H^{1/2}(\Gamma) \\ \text{such that } \|\mu_h - v\|_{H^{1/2}(\Gamma)} \rightarrow 0 \quad \text{as } h \rightarrow 0$$

Again, we introduce a family of projections $Q_h: H^{1/2}(\Gamma) \rightarrow \mathcal{B}_h$ that satisfies the uniform boundedness condition

$$\|Q_h\| \leq c_p \quad \text{for all } 0 < h \leq h_0 \quad (119)$$

For the Galerkin method to (85) with given $\sigma \in H_0^{1/2}(\Gamma)$ (hence $\omega = 0$), we have two variational formulations, (81) and (82). In the latter case, we can use the full finite-dimensional space \mathcal{B}_h , and the Galerkin equations read

Find $\mu_h \in \mathcal{B}_h$ satisfying

$$\tilde{a}_{2c}(\mu_h, \mu_h) := a_{2c}(\mu_h, \mu_h) + (\mu_h, 1)(1, \mu_h) = (\mu_h, \sigma) \\ \text{for all } v_h \in \mathcal{B}_h \quad (120)$$

Since $\tilde{a}_{2c}(\mu, \mu)$ is $H^{1/2}(\Gamma)$ -elliptic, we now obtain the results that are completely analogous to those in Theorem 9, which will also be summarized below.

One may also incorporate the side condition $(1, \mu) = 0$ into the family of finite-dimensional subspaces

$$\mathcal{B}_{0,h} := \{\mu_h \in \mathcal{B}_h \mid (\mu_h, 1) = 0\} \subset H_0^{1/2}(\Gamma)$$

and execute the Galerkin method for (120) or (127) on the subspace $\mathcal{B}_{0,h} \subset H^{1/2}(\Gamma)$. Again, we obtain Céa's estimate since a_{2c} is $H_0^{1/2}(\Gamma)$ -elliptic.

Now we consider a_{2c} and suppose for a moment that the composed operator DK is on Γ available in explicit form. Then the Galerkin equations for (69) read

Find $\mu_h \in \mathcal{B}_{0,h} \subset H_0^{1/2}(\Gamma)$ from

$$a_{2c}(\mu_h, \mu_h) = (\mu_h, (B - C)\mu_h) = (\mu_h, (B - C)\mu) \\ \text{for all } v_h \in \mathcal{B}_{0,h} \quad (121)$$

where $\mu \in H^{1/2}(\Gamma)$ is the unique solution of (69) in the subspace $H_0^{1/2}(\Gamma)$. Here, B is the $H_0^{1/2}(\Gamma)$ -elliptic operator part of the bilinear form (67) satisfying

$$(Bv, v) \geq \alpha_2 \|v\|_{H^{1/2}(\Gamma)}^2 \quad \text{for all } v \in H_0^{1/2}(\Gamma) \text{ with } \alpha_2 > 0 \\ \text{and } C: H_0^{1/2}(\Gamma) \rightarrow H_0^{-1/2}(\Gamma) \text{ is compact. Hence, (121) can also be written as}$$

$$Q_h^* B Q_h [I - (Q_h^* B Q_h)^{-1} Q_h^* C] \mu_h = Q_h^* (B - C) \mu$$

With the Galerkin projection $G_{hB} := (Q_h^* B Q_h)^{-1} Q_h^* B$ of B , which is uniformly bounded because of (122), and with the operators

$$L := I - B^{-1}C \quad \text{and} \quad L_h := I - G_{hB} B^{-1}C$$

we find the equation

$$L_h \mu_h = G_{hB} L \mu$$

Since $\lim_{h \rightarrow 0} \|(G_{hB} - I)v\|_{H^{1/2}(\Gamma)} = 0$ for every $v \in H_0^{1/2}(\Gamma)$ and since $B^{-1}C: H_0^{1/2}(\Gamma) \rightarrow H^{-1/2}(\Gamma)$ is compact, we obtain

$$\|L_h - L\| = \|(I - G_{hB})B^{-1}C\| \rightarrow 0 \quad \text{as } h \rightarrow 0$$

with respect to the operator norm $\|\cdot\|$.

Moreover, $L^{-1} = (B - C)^{-1}B$ exists; hence, there are constants $\alpha_0 > 0$ and c_0 such that L^{-1} exists and

$$\|L_h^{-1}\| \leq c_0 \quad \text{for all } 0 < h \leq h_0$$

(see Proposition 1.8 and Theorem 1.11 in Anselone (1971) and Kantorovich and Akilov (1964)). Therefore,

$$\mu_h = L_h^{-1} G_{hB} L \mu =: G_{h(B-C)} \mu$$

and the Galerkin projection $G_{h(B-C)}$ to $B - C$ in (121) is uniformly bounded on $H_0^{1/2}(\Gamma)$. Therefore, Céa's estimate holds for (121) as well as for the stabilized version (82).

In practice, however, the Galerkin weights

$$a_{2c}(\mu_j, \mu_k) = (D\mu_j, (\frac{1}{2}I - K)\mu_k)$$

belonging to a basis $(\mu_j)_{j=1}^{N_h}$ can only be computed accurately enough if the outer numerical integrations are sufficiently accurate or if an intermediate space \mathcal{S}_h is used in combination with the duality as in Steinbach (2002). Here, we circumvent these difficulties by replacing (121) by the following system of Galerkin equations, which, in fact, correspond to some preconditioning.

For given $\varphi \in H_0^{1/2}(\Gamma)$, compute $\varphi_h \in \mathcal{B}_{0,h}$ from

$$a_{2c}(\varphi_h, \varphi_h) = (\varphi_h, D\varphi) \quad \text{for all } v_h \in \mathcal{B}_{0,h} \quad (123)$$

With φ_h obtained, compute $\mu_h \in \mathcal{B}_{0,h}$ from

$$(\varphi_h, (\frac{1}{2}I - K)\mu_h) = (\varphi_h, \varphi_h) \quad \text{for all } v_h \in \mathcal{B}_{0,h} \quad (124)$$

For (123), we have a unique solution satisfying

$$\|\varphi_h\|_{H^{1/2}(\Gamma)} \leq c' \|\varphi\|_{H^{1/2}(\Gamma)} \leq c \|\mu\|_{H^{1/2}(\Gamma)}$$

where $\mu \in H_0^{1/2}(\Gamma)$ is the unique solution to (69). In fact, equation (124) is equivalent to the integral equation of the second kind

$$\mu_h = (\frac{1}{2}I + K)\mu_h + \varphi_h \quad \text{in } H_0^{1/2}(\Gamma) \subset H^{1/2}(\Gamma)$$

and Theorem 5 implies unique solvability. Moreover, we obtain for the Galerkin projection $\mu_h := G_{hD}\varphi$ associated with the system (123), (124) the uniform estimate

$$\|G_{hD}\mu\|_{H^{1/2}(\Gamma)} = \|\mu_h\|_{H^{1/2}(\Gamma)} \leq \|\mu_h\|_{ED} \\ \leq \frac{c}{1 - c_K} \|\varphi_h\|_{ED} \leq c_2 \|\varphi\|_{H^{1/2}(\Gamma)} \leq c_2 \|\mu\|_{H^{1/2}(\Gamma)}$$

where c_2 is independent of h . Consequently, Céa's estimate is also valid for the system (123), (124), where only the Galerkin weights of D and K are needed.

In rather the same manner, the Galerkin methods can be employed and analyzed for the Fredholm integral equation of the second kind (75) and its variational version (78).

For (81) respectively (85), with the additional unknown $\omega \in \mathbb{R}$, we now define the bilinear form

$$a_{2c}^1((v, \kappa); (\mu, \omega)) := a_{2c}(v, \mu) + \omega(v, 1) + (1, \mu)\kappa \quad (125)$$

on the augmented space $\mathcal{H} := (H^{1/2}(\Gamma) \times \mathbb{R})$, where a_{2c}^1 satisfies Gårding's inequality

$$a_{2c}^1((v, \kappa); (v, \kappa)) + c((v, \kappa); (v, \kappa)) \\ \geq \alpha_2^1(\|v\|_{H^{1/2}(\Gamma)}^2 + |\kappa|^2) \quad \text{on } \mathcal{H} = (H^{1/2}(\Gamma) \times \mathbb{R}) \quad (126)$$

where $c((v, \kappa); (v, \kappa)) = 2\kappa\omega + 2(v, 1)(1, \mu)$ is a compact bilinear form. Then the Galerkin equations for a_{2c}^1 corresponding to (85) now read

Find $(\mu_h, \omega_h) \in \mathcal{H}_h := (\mathcal{B}_h \times \mathbb{R})$ satisfying

$$a_{2c}^1((v_h, \kappa); (\mu_h, \omega_h)) \\ = a_{2c}(v_h, \mu_h) + \omega_h(v_h, 1) + (1, \mu_h)\kappa = (v_h, \sigma) \quad (127)$$

for all $(v_h, \kappa) \in \mathcal{H}_h := (\mathcal{B}_h \times \mathbb{R})$.

Since the solution to (85) is unique in \mathcal{H} , from Gårding's inequality (126), it follows in the same manner as before for a_{2c} that the Galerkin projection $\mathcal{G}_h: \mathcal{H} \rightarrow \mathcal{H}_h$ is uniformly bounded and Céa's estimate is again valid.

We summarize all these results in the following theorem.

Theorem 10. Let $\mu_h \in \mathcal{B}_h$ be the unique solution of (120) or $\mu_h \in \mathcal{B}_{0,h}$ the unique solution of (121) or of (123), (124), and let $(\mu_h, \omega_h) \in \mathcal{H}_h := (\mathcal{B}_h \times \mathbb{R})$ be the unique solution pair of (127). Then, we have Céa's estimates

$$\|\mu - \mu_h\|_{H^{1/2}(\Gamma)} \leq \tilde{c} \inf_{v_h \in \mathcal{B}_h} \|\mu - v_h\|_{H^{1/2}(\Gamma)}$$

and

$$\|(\mu, \omega) - (\mu_h, \omega_h)\|_{\mathcal{H}} \leq c_1 \inf_{(v_h, \kappa) \in \mathcal{H}_h} \|(\mu, \omega) - (v_h, \kappa)\|_{\mathcal{H}} \\ = c_1 \inf_{v_h \in \mathcal{B}_h} \|\mu - v_h\|_{H^{1/2}(\Gamma)}$$

with some constants \tilde{c}, c_1 , independent of $\mathcal{B}_h, \mathcal{B}_{0,h}$ or \mathcal{H}_h , respectively.

Clearly, the corresponding Galerkin projections are uniformly bounded:

$\tilde{\mathcal{G}}_{hD}: H^{1/2}(\Gamma) \rightarrow \mathcal{B}_h$ defined by (120) satisfies

$$\|\tilde{\mathcal{G}}_{hD}\| \leq c$$

$G_{hD}: H_0^{1/2}(\Gamma) \rightarrow \mathcal{B}_{0,h}$ defined by (121), or by

$$(123), (124), \text{ satisfies } \|G_{hD}\| \leq c$$

and

$$\mathcal{G}_h: \mathcal{H} \rightarrow \mathcal{H}_h \text{ defined by (127) satisfies} \\ \|\mathcal{G}_h\| \leq c$$

4.2 Optimal order of convergence

The estimates in Theorems 9 and 10 are basic abstract error estimates that provide sufficient conditions for the convergence of Galerkin solutions, if the family \mathcal{H}_h of subspaces of the energy space \mathcal{H} have the approximation property (ap)

$$\lim_{h \rightarrow 0} \inf_{(v_h, \kappa_h) \in \mathcal{H}_h} \|(v_h, \kappa_h) - (u, \omega)\|_{\mathcal{H}} = 0$$

However, in order to know the precise order of convergence, one needs more specific properties from the approximation theory concerning the approximate subspaces and regularity results for the exact solutions to the BIEs under consideration.

We first need the notion of the higher-order Sobolev spaces other than those introduced in the beginning of this section.

(1) *Sobolev space* $H^m(\Omega)$, $m \in \mathbb{N}_0$. We start with the function space

$$C_m^\infty(\Omega) := \{u \in C^\infty(\Omega) \mid \|u\|_{H^m(\Omega)} < \infty\}$$

with

$$\|u\|_{H^m(\Omega)} := \left\{ \sum_{|\alpha| \leq m} \int_{\Omega} |\partial^\alpha u|^2 dx \right\}^{1/2} \\ \partial^\alpha := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}$$

where $\alpha \in \mathbb{N}_0^n$ and $|\alpha| = \alpha_1 + \cdots + \alpha_n$. Then, $H^m(\Omega)$, the Sobolev space of order m , is defined as the completion of $C_m^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{H^m(\Omega)}$. By this we mean that for every $u \in H^m(\Omega)$ there exists a sequence $\{u_k\}_{k \in \mathbb{N}} \subset C_m^\infty(\Omega)$ such that

$$\lim_{k \rightarrow \infty} \|u - u_k\|_{H^m(\Omega)} = 0 \quad (128)$$

We recall that two Cauchy sequences $\{u_k\}$ and $\{v_k\}$ in $C_m^\infty(\Omega)$ are said to be equivalent if and only if $\lim_{k \rightarrow \infty} \|u_k - v_k\|_{H^m(\Omega)} = 0$. This implies that $H^m(\Omega)$, in fact, consists of all equivalence classes of Cauchy sequences and that the limit u in (128) is just a representative for the class of

equivalent Cauchy sequences $\{u_k\}$. The space $H^m(\Omega)$ is a Hilbert space with the inner product defined by

$$(u, v)_{H^m(\Omega)} := \sum_{|\alpha| \leq m} \int_{\Omega} \partial^\alpha u \overline{\partial^\alpha v} dx$$

(2) *Sobolev space* $H^s(\Omega)$ for $0 < s \in \mathbb{R}$. By setting

$$s = m + \lambda \quad \text{with } m \in \mathbb{N}_0 \quad \text{and } 0 < \lambda < 1$$

the Sobolev space $H^s(\Omega)$ of order s is defined as the completion of

$$C_s^\infty(\Omega) := \{u \in C^{m+1}(\Omega) \mid \|u\|_{H^m(\Omega)} < \infty\}$$

with respect to the norm

$$\|u\|_{H^s(\Omega)} := \left\{ \|u\|_{H^m(\Omega)}^2 + \sum_{|\alpha|=m} \int_{\Omega} \int_{\Omega} \frac{|\partial^\alpha u(x) - \partial^\alpha u(y)|^2}{|x - y|^{n+2\lambda}} dx dy \right\}^{1/2} \quad (129)$$

the so-called *Slobodetskii norm*. Note that the second part in the definition (129) of the norm $\|\cdot\|_{H^s(\Omega)}$ gives the L^2 -version of fractional differentiability, which is compatible to the pointwise version in $C^{m+\lambda}(\Omega)$, the Hölder m -continuously differentiable function space (cf. Section 2.2). The space $H^s(\Omega)$ is again a Hilbert space with the inner product defined by

$$(u, v)_{H^s(\Omega)} := (u, v)_{H^m(\Omega)} + \sum_{|\alpha|=m} \int_{\Omega} \int_{\Omega} \frac{(\partial^\alpha u(x) - \partial^\alpha u(y))(\partial^\alpha v(x) - \partial^\alpha v(y))}{|x - y|^{n+2\lambda}} dx dy$$

We remark that the Sobolev spaces $H^m(\Omega)$ and $H^s(\Omega)$ can also be defined by the theory of distributions (see e.g. Adams, 1975). Owing to Meyers and Serrin (1964), these two approaches will lead to spaces that are equivalent. We believe that the definition of Sobolev spaces based on the completion is more intuitive.

(3) *Sobolev spaces* $H^s(\Gamma)$ and $H^{-t}(\Gamma)$ for $0 \leq s \in \mathbb{R}$. For $s = 0$, we set $H^0(\Gamma) := L^2(\Gamma)$. For $s > 0$, the simplest way to define the trace spaces on Γ is to use extensions of functions defined on Γ to functions in Sobolev spaces defined in Ω . More precisely, let

$$C_s^0(\Gamma) := \{u \in C^0(\Gamma) \mid \text{to } u \text{ there exists } \tilde{u} \in H^{s+1/2}(\Omega) \\ \text{such that } \gamma_0 \tilde{u} := \tilde{u}|_\Gamma = u \text{ on } \Gamma\}$$

Then the *natural trace space* $H^s(\Gamma)$ is defined by

$$H^s(\Gamma) := \overline{C_s^0(\Gamma)}^{\|\cdot\|_{H^s(\Gamma)}}$$

the completion of $C_s^0(\Gamma)$ with respect to the norm

$$\|u\|_{H^s(\Gamma)} := \inf_{\gamma_0 \tilde{u} = u} \|\tilde{u}\|_{H^{s+1/2}(\Omega)} \quad (130)$$

We note that with this definition the inner product cannot be deduced from (130), although it can be shown that $H^s(\Gamma)$ is a Hilbert space. On the other hand, the trace theorem holds by definition, namely,

$$\|\gamma_0 \tilde{u}\|_{H^s(\Gamma)} \leq \|\tilde{u}\|_{H^{s+1/2}(\Omega)} \quad \text{for every } \tilde{u} \in H^{s+1/2}(\Omega)$$

For $s < 0$, we can define the space $H^s(\Gamma)$ as the dual of $H^{-s}(\Gamma)$ with respect to the $L^2(\Gamma)$ -duality (\cdot, \cdot) ; that is, the completion of $L^2(\Gamma)$ with respect to the norm

$$\|u\|_{H^{-s}(\Gamma)} := \sup_{\{v, u\}_{L^2(\Gamma)} = 1} |(v, u)| \quad (131)$$

These are the boundary spaces of negative orders. On the other hand, the boundary Γ is usually identified with \mathbb{R}^{n-1} by means of local parametric representations of the boundary, and hence the trace spaces are defined to be isomorphic to the Sobolev space $H^s(\mathbb{R}^{n-1})$. This means that the spaces $H^s(\Gamma)$ behave like the spaces $H^s(\mathbb{R}^{n-1})$. For further discussion of this latter approach, see, for example, Nečas (1967) and Aubin (1972). However, for our purpose, the above definitions of trace spaces based on (130) and (131) will be sufficient.

We are now in a position to specify the approximation spaces \mathcal{S}_h by using boundary elements. In particular, we assume that $\mathcal{S}_h = \mathcal{S}_h^{e,m}$ is a regular boundary element space as introduced in Babuška and Aziz (1977). That is, $\mathcal{S}_h^{e,m}$ with $\ell, m \in \mathbb{N}_0$ and $m+1 \leq \ell$ has the following properties:

Definition 7 (Approximation property) Let $t \leq s \leq \ell$ and $t < m + (1/2)$ for $n = 2$ or $t \leq m$ for $n = 3$. Then, there exists a constant c such that for any $v \in H^s(\Gamma)$ a sequence $\chi_h \in \mathcal{S}_h^{e,m}$ exists and satisfies the estimate

$$\|v - \chi_h\|_{H^t(\Gamma)} \leq ch^{t-t} \|v\|_{H^s(\Gamma)} \quad (132)$$

Definition 8 (Inverse property) For $t \leq s < m + (1/2)$ for $n = 2$ or $t \leq s \leq m$ for $n = 3$, there exists a constant M such that for all $\chi_h \in \mathcal{S}_h^{e,m}$,

$$\|\chi_h\|_{H^t(\Gamma)} \leq Mh^{t-t} \|\chi_h\|_{H^s(\Gamma)} \quad (133)$$

It can be shown that the approximation and inverse properties together imply for the L^2 -projections P_h and

Q_h the respective uniform bounds (116) and (119). For nonregular grids, in general, the inverse property will not be true anymore in this form, but might be replaced by (116) and (119), respectively. For the Galerkin solution λ_h of (111) in $\mathcal{S}_h^{e,m}$ with the properties (132) and (133), we then have the following error estimate.

Theorem 11. For $-1/2 \leq t \leq s \leq \ell$, $t \leq m$, we have the asymptotic error estimate of optimal order

$$\|\lambda - \lambda_h\|_{H^t(\Gamma)} \leq ch^{s-t} \|\lambda\|_{H^s(\Gamma)} \quad (134)$$

for the Galerkin solution λ_h of (111), provided the exact solution $\lambda \in H^s(\Gamma)$.

Proof. For $t = -1/2$, the estimate (134) follows immediately from Céa's lemma (114) and the approximation property (132). That is,

$$\|\lambda - \lambda_h\|_{H^{-1/2}(\Gamma)} \leq ch^{s+1/2} \|\lambda\|_{H^s(\Gamma)} \quad (135)$$

For $t \geq -1/2$, we need the inverse property (133) in addition to the approximation property (132) and the established estimate (135):

$$\begin{aligned} \|\lambda - \lambda_h\|_{H^t(\Gamma)} &\leq \|\lambda - \chi_h\|_{H^t(\Gamma)} + \|\chi_h - \lambda_h\|_{H^t(\Gamma)} \\ &\quad \text{with } \chi_h \text{ for } \lambda_h \text{ as in (132)} \\ &\leq ch^{s-t} \|\lambda\|_{H^s(\Gamma)} + Mh^{-1/2-t} \|\lambda_h - \lambda_h\|_{H^{-1/2}(\Gamma)} \\ &\quad \text{from (132) and (133)} \\ &\leq ch^{s-t} \|\lambda\|_{H^s(\Gamma)} + Mh^{-1/2-t} \\ &\quad \times (\|\chi_h - \lambda\|_{H^{-1/2}(\Gamma)} + \|\lambda - \lambda_h\|_{H^{-1/2}(\Gamma)}) \\ &\leq c'h^{s-t} \|\lambda\|_{H^s(\Gamma)} \end{aligned}$$

The last step follows from the approximation property (132) and the estimate (135) in the energy space $H^{-1/2}(\Gamma)$. \square

For the estimate of the Galerkin solution pair (u_h, ω_h) of the equation (127), we note that the approach will be exactly the same if we modify the approximation property (132) and the inverse property (133), respectively, as

$$\|v - \chi_h\|_{H^t(\Gamma)} + |\omega - \kappa_h| \leq ch^{t-t} \|v\|_{H^s(\Gamma)} \quad (136)$$

$$\|\chi_h\|_{H^t(\Gamma)} + |\kappa_h| \leq M'h^{-t} \{\|\chi_h\|_{H^s(\Gamma)} + |\kappa_h|\} \quad (137)$$

with $M' := \max\{1, M\}$, where M is the constant in (133). We have the estimate in the energy space $(\mathcal{H} = H^{1/2}(\Gamma) \times \mathbb{R})$, in this case) from Céa's lemma (Theorem 10) by making use of the approximation property. Then for estimates

in stronger norms, the inverse property will be used. We summarize our results for the Galerkin solution of (127) in the following theorem.

Theorem 12. Let $\mathcal{H}_h := S_h^{t,m} \times \mathbb{R}$ have the approximation property (136) and the inverse property (137). Then the Galerkin solution pair (μ_h, ω_h) for the equation (127) converges for $1/2 \leq t \leq s \leq \ell$ and $t \leq m$ asymptotically as

$$\|\mu - \mu_h\|_{H^t(\Gamma)} + |\omega - \omega_h| \leq ch^{t-1} \left\{ \|\mu\|_{H^t(\Gamma)} + |\omega| \right\} \quad (138)$$

4.3 Aubin–Nitsche lemma

The estimates (134) and (138) are optimal with respect to the $H^t(\Gamma)$ -norm for $\alpha \leq t$, where $H^\alpha(\Gamma)$ with $\alpha = -1/2$ or $1/2$ are the respective energy spaces. In order to obtain an optimal rate of convergence for $t < \alpha$ also, we need the so-called Aubin–Nitsche lemma for BIEs; see Hsiao and Wendland (1981a) and Costabel and Stephan (1988). The Aubin–Nitsche lemma is often referred to as the *Nitsche's trick* when one uses FEMs to solve BVPs for PDEs. By applying this lemma, one also obtains the so-called superapproximation of the approximate solution; see Nitsche (1968) and also Ciarlet (1978). The proof of the Aubin–Nitsche lemma is based on duality arguments. However, for this purpose, one needs some additional mapping properties of the boundary integral operators involved.

To be more specific, let us return to the Galerkin equations (111) associated with the simple-layer boundary integral operator V defined by (60). What we need is the following mapping property in addition to the Gårding inequality (103). The mapping

$$V: H^t(\Gamma) \rightarrow H^{t+1}(\Gamma) \quad (139)$$

is a continuous isomorphism for any $t \in \mathbb{R}$. By an isomorphism, we mean that V is one-to-one and onto. Hence, by the continuity of V , the inverse V^{-1} is also continuous. We have shown (139) for $t = -1/2$. In fact, this property is true and has been shown in Hsiao and Wendland (1977) in the plane and can be easily established by using the theory of pseudodifferential operators; see, for example, Hsiao and Wendland (2005). Now, under the condition (139), because $V^* = V$ is also a continuous isomorphism of the same order (see Taylor, 1981), that is,

$$V = V^*: H^{-t-1}(\Gamma) \rightarrow H^{-t}(\Gamma)$$

is a continuous isomorphism for any $t \in \mathbb{R}$ (see Figure 2).

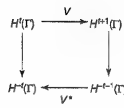


Figure 2. Continuous isomorphisms.

Hence, in particular, we have the estimate

$$\|\chi\|_{H^{-t-1}(\Gamma)} \leq c \|V^* \chi\|_{H^{-t}(\Gamma)} \quad \text{for every } t \in \mathbb{R} \quad (140)$$

Now let $e_h := \lambda - \lambda_h$ denote the error of the Galerkin solution of (111). Then for $\chi \in H^{-t-1}(\Gamma)$ with $t \leq -1/2$, we have

$$\begin{aligned} |(e_h, V^* \chi)| &= |(V e_h, \chi)| = |(V e_h, \chi - \chi_h) + (V e_h, \chi_h)| \\ &= |(V e_h, \chi - \chi_h)| \end{aligned}$$

following from the orthogonality of e_h with respect to $a_{12}(\cdot, \cdot)$. The latter implies that

$$\begin{aligned} |(e_h, V^* \chi)| &\leq \inf_{\chi_h \in S_h^{t,m}} |(V e_h, \chi - \chi_h)| \\ &\leq c \|e_h\|_{H^{-t/2}(\Gamma)} \inf_{\chi_h \in S_h^{t,m}} \|\chi - \chi_h\|_{H^{-t/2}(\Gamma)} \\ &\leq c' \|e_h\|_{H^{-t/2}(\Gamma)} h^{-t-1/2} \|\chi\|_{H^{-t-1}(\Gamma)} \end{aligned}$$

for $-t-1 \leq \ell$ from the approximation property (132). The estimates (135) and (140) imply that

$$|(e_h, V \chi)| \leq ch^{t-1} \|\lambda\|_{H^t(\Gamma)} \|V \chi\|_{H^{-t}(\Gamma)}$$

from which the estimate

$$\|e_h\|_{H^t(\Gamma)} = \sup_{|v|_{H^{-t}(\Gamma)} \leq 1} |(e_h, v)| \leq ch^{t-1} \|\lambda\|_{H^t(\Gamma)}$$

follows for $-1-\ell \leq t \leq -1/2$. Thus, we have proved the following lemma.

Lemma 7. The error estimate (134) remains valid also for $-1-\ell \leq t \leq -1/2$, where we do not need the inverse property.

Now, for the equation (127), in order to extend the results of (138) for $t < 1/2$, we need the regularity results for the operator A defined by the sesquilinear form $a_{\mathcal{H}}$ in (125) and an estimate similar to (140), but for the operator A^*

adjoint to A , namely, the estimate

$$\|(v, \kappa)\|_{H^{-t+1}(\Gamma) \times \mathbb{R}} \leq \|A^*(v, \kappa)\|_{H^{-t}(\Gamma) \times \mathbb{R}} \quad \text{for } t \in \mathbb{R} \quad (141)$$

These results are indeed known, and for a proof, see Hsiao and Wendland (2005). The corresponding Lemma 7 now reads:

Lemma 8. The asymptotic error estimate in Theorem 12 can be extended for $-1-\ell \leq t \leq 1/2 \leq s$ and $1/2 \leq m$ with the additional regularity condition (141).

The proof of Lemma 8 proceeds identically as that for Lemma 7.

One would like to push the error estimates in Sobolev spaces of lower order as far as one can. The purpose for doing so is that in this way one may obtain some kind of superconvergence. For instance, if we substitute the Galerkin solution λ_h of (111) into the boundary potential in (57),

$$u_h(x) := \int_{\Gamma} E(x, y) \lambda_h(y) dy, \quad \text{for } x \in \Omega$$

then both $u := V\lambda$ and $u_h = V\lambda_h$ satisfy the PDE in (56). Moreover, we see that for any compact subset Ω' of Ω , we have

$$\begin{aligned} |u(x) - u_h(x)| &= |(\lambda - \lambda_h, E(x, \cdot))| \\ &\leq \|\lambda - \lambda_h\|_{H^{-t}(\Gamma)} \|E(x, \cdot)\|_{H^t(\Gamma)} \\ &\leq c(d) h^{t+1/2} \|\lambda\|_{H^t(\Gamma)} \quad \text{for } x \in \Omega' \subseteq \Omega \end{aligned}$$

where $c(d)$ is a constant depending on $d := \sup\{|x - y| \text{ for } x \in \Omega' \text{ and } y \in \Gamma\}$. By the same approach, it is easy to see that similar estimates hold for the derivatives as well. That is, we have the superconvergence results

$$|D^{\beta} u(x) - D^{\beta} u_h(x)| = O(h^{2t-2\beta}), \quad x \in \Omega' \subseteq \Omega$$

for $\alpha = -1/2$ and $\alpha = 1/2$, respectively for the corresponding boundary potentials u in terms of λ and μ and their Galerkin approximations u_h (see e.g. Hsiao and Wendland, 1981b).

4.4 Stability and Ill-posedness

As is well known, an integral equation of the first kind such as (58) is not well posed in the sense that small L^2 disturbance in the data φ may produce arbitrarily large discrepancies in the solutions. On the other hand, ill-posed problems are frequently treated by the regularization method as in Tikhonov and Arsenin (1977) and Natterer

(1986). In order to see the connections, we allow the possibility of imprecise data and replace φ in (56) by its perturbation φ_ε . We assume that

$$\|\varphi - \varphi_\varepsilon\|_{L^2(\Gamma)} \leq \varepsilon \quad (142)$$

holds, where ε is a small parameter. Now we denote by λ_h^* the corresponding Galerkin solution of (111) with φ replaced by φ_ε . Then, it follows from (117) that

$$\lambda - \lambda_h^* = (\lambda - \lambda_h) + (P_h^* V P_h^*)^{-1} P_h^* (\varphi - \varphi_\varepsilon)$$

where P_h^* is now the L^2 -projection satisfying (116). Consequently, we obtain the estimate

$$\begin{aligned} \|\lambda - \lambda_h^*\|_{H^{-t/2}(\Gamma)} &\leq \|(\lambda - \lambda_h)\|_{H^{-t/2}(\Gamma)} + c \|P_h^* (\varphi - \varphi_\varepsilon)\|_{H^{-t/2}(\Gamma)} \end{aligned}$$

from the uniform boundedness of $(P_h^* V P_h^*)^{-1}$. We have obtained an estimate for the first term on the right-hand side. In order to use the information of (142), we need to employ the inverse property (133) to dominate the second term on the right-hand side by the L^2 -norm. This leads to the abstract error estimate in the form of (114), that is,

$$\|\lambda - \lambda_h^*\|_{H^{-t/2}(\Gamma)} \leq c \left\{ \inf_{\chi \in S_h^{t,m}} \|\lambda - \chi\|_{H^{-t/2}(\Gamma)} + h^{-1/2} \varepsilon \right\}$$

from which general estimates can be obtained as before. Of particular interest is the L^2 estimate

$$\|\lambda - \lambda_h^*\|_{L^2(\Gamma)} \leq c(h^t \|\lambda\|_{H^t(\Gamma)} + h^{-1/2} \varepsilon) \quad (143)$$

for $\lambda \in H^t(\Gamma)$. The estimate (143) provides us some guidance concerning the choice of the mesh size h in numerical computations (see Hsiao, 1986). From (143), it is easy to see that for given ε , there is an optimal choice of h ,

$$h_{\text{opt}} = \varepsilon^{1/(t+1)}$$

With this choice of h , we have

$$\|\lambda - \lambda_h^*\|_{L^2(\Gamma)} = O(\varepsilon^{t/(t+1)}) \quad \text{as } \varepsilon \rightarrow 0^+$$

which coincides with the result obtained by the Tikhonov-regularization method as in Tikhonov and Arsenin (1977) and Natterer (1986), if the regularization parameter there is chosen optimally. Hence, for this type of problems, the discretization via Galerkin's method is already an optimal regularization.

It is also known that the L^2 -condition number of the Galerkin equation (111) is unbounded. This can be seen as

follows. Following Hsiao and Wendland (1977), Wendland (1983, Corollary 2.9), and Hsiao (1987), we may write from (118),

$$\lambda_h = (\mathbb{P}_h^* V \mathbb{P}_h)^{-1} (\mathbb{P}_h^* V \lambda) = (\mathbb{P}_h^* V \mathbb{P}_h)^{-1} (\mathbb{P}_h^* \varphi) = G_{hV} V^{-1} (\mathbb{P}_h^* \varphi)$$

where \mathbb{P}_h and G_{hV} are the projection and the Galerkin operator respectively. Applying Theorem 11 with λ replaced by $V^{-1}(\mathbb{P}_h^* \varphi)$ and $s = t = 0$, we obtain the estimate

$$\begin{aligned} \|V^{-1}(\mathbb{P}_h^* \varphi) - \lambda_h\|_{L^2(\Gamma)} &\leq c \|V^{-1}(\mathbb{P}_h^* \varphi)\|_{L^2(\Gamma)} \\ &\leq c \|\mathbb{P}_h^* \varphi\|_{H^1(\Gamma)} \leq c M h^{-1} \|\mathbb{P}_h \varphi\|_{L^2(\Gamma)} \end{aligned}$$

from the inverse property (133). Then we have

$$\begin{aligned} \|\lambda_h\|_{L^2(\Gamma)} &\leq \|V^{-1}(\mathbb{P}_h^* \varphi) - \lambda_h\|_{L^2(\Gamma)} + \|V^{-1}(\mathbb{P}_h^* \varphi)\|_{L^2(\Gamma)} \\ &\leq c' h^{-1} \|\mathbb{P}_h \varphi\|_{L^2(\Gamma)} \end{aligned}$$

for some constant c' , independent of h . The latter implies that

$$\|(\Lambda_h)^{-1}\| \leq c' h^{-1}$$

where $\Lambda_h := \mathbb{P}_h^* V \mathbb{P}_h$ denotes the coefficient matrix of the Galerkin equation (111) and $\|\cdot\|$ is the spectral norm here. On the other hand, we have from the continuity of V

$$\|\mathbb{P}_h^* V \mathbb{P}_h \lambda_h\|_{L^2(\Gamma)} \leq c \|\lambda_h\|_{H^{-1}(\Gamma)} \leq c \|\lambda_h\|_{L^2(\Gamma)}$$

which implies that

$$\|\Lambda_h\| \leq c$$

Hence the L^2 -condition number satisfies

$$\text{cond}_{L^2}(\Lambda_h) := \|\Lambda_h\| \|\Lambda_h^{-1}\| = O(h^{-1})$$

Similarly, we arrive at the estimates for the Galerkin solution of (127):

$$\begin{aligned} \|(\mu_h, \omega_h)\|_{L^2(\Gamma) \times \mathbb{R}} &\leq (c_1 h + c_2) \|\Lambda_h(\mu_h, \omega_h)\|_{L^2(\Gamma) \times \mathbb{R}} \\ \|\Lambda_h(\mu_h, \omega_h)\|_{L^2(\Gamma) \times \mathbb{R}} &\leq c h^{-1} \|(\mu_h, \omega_h)\|_{L^2(\Gamma) \times \mathbb{R}} \end{aligned}$$

where Λ_h now denotes the stiffness matrix of the Galerkin equations of (127). This gives the estimate for the condition number of the Galerkin equations for (125) and, in the same manner, also for the Galerkin equations to (120) to (82), (121) or (123), (124) to (69) and those to (78) and (86); that is,

$$\text{cond}_{L^2(\Gamma)}(\Lambda_h) = O(h^{-1})$$

In all these cases, we see that the condition numbers will be unbounded as $h \rightarrow 0^+$. This is not surprising. We note that although (59) is a BIE of the second kind, its variational formulation (84) with the help of the operator D actually corresponds to the weak formulation of a BIE of the first kind, according to our classification in the next section. As will be seen, the results presented here for (58) and for (59) (or rather for its weak form (84)) are just special cases of a general result for BIEs of the first kind. We will return to this discussion in the next section.

However, if Γ is sufficiently smooth, as, for example, in C^2 (or at least Lyapunov), then, in case of the Laplacian or the Lamé system, the integral operators of the second kind in (59) and in (75) are bounded in $L^2(\Gamma)$ and satisfy Gårding inequalities there of the form

$$\text{Re}(\mu, (\tfrac{1}{2}I \pm K)\mu + C_s \mu)_{L^2(\Gamma)} \geq \alpha_0 \|\mu\|_{L^2(\Gamma)}^2 \quad (144)$$

for all $\mu \in L^2(\Gamma)$. Then one may apply the Galerkin method for (59) or (75) directly in $L^2(\Gamma)$, and the corresponding condition numbers will be uniformly bounded on appropriate boundary element spaces, which is very advantageous for fast iterative solution methods. Therefore, based on the identities (24), one tries to find preconditioning operators for the BIEs of the first kind, as in (34) or (46), to be converted into integral equations of the second kind (see Steinbach and Wendland, 1998). Unfortunately, for general Lipschitz boundaries Γ , inequalities of the type (144) are not known and might not even be true anymore.

5 THE ROLE OF SOBOLEV INDEX

This section is the heart of the chapter. After presenting the concrete examples in the previous sections, it is the purpose of this section to discuss the BEM from a more general point of view. Fundamental concepts and underlying mathematical principles, which may have already appeared in different forms in the previous special model problems, will be addressed once more. General results, which contain those obtained in the previous examples as special cases, for the Galerkin-BEM will be collected in this section.

5.1 Order and classifications

We begin with the classification of BIEs based on the mapping properties of the BIE operators involved. Let us consider a BIE of the general form

$$A\sigma = f \quad \text{on } \Gamma \quad (145)$$

Here, A is the operator on Γ and σ is the unknown boundary charge (unknown density or moment function), and f is the given data function on the boundary Γ . We assume that $f \in H^{s-\alpha}(\Gamma)$, $s \in \mathbb{R}$, where 2α is a fixed constant. (It is assumed that the boundary manifold Γ is sufficiently smooth for the corresponding s and α to be specified.)

Definition 9. We say that the order of the BIE operator A is 2α if the mapping

$$A : H^{s+\alpha}(\Gamma) \mapsto H^{s-\alpha}(\Gamma)$$

for any $s \in \mathbb{R}$ with $|s| \leq s_0$ is continuous, where s_0 is some fixed positive constant.

For example, if the direct method is used for solving the interior and exterior Dirichlet and Neumann problems for the Lamé equations (8) and (11), we will arrive at the following typical boundary integral operators:

$$\begin{aligned} 2\alpha &= -1, & A &= V \\ 2\alpha &= 0, & A &= \tfrac{1}{2}I \pm K \quad \text{or } A = \tfrac{1}{2}I \mp K' \\ 2\alpha &= +1, & A &= D \end{aligned}$$

Here, V , K , K' , and D are the four basic BIE operators introduced in (20), (21), (22), and (23) respectively. Boundary integral equations such as (145) are classified according to the order 2α of the boundary integral operator A .

We call (145) a first-kind BIE if $2\alpha < 0$. If $2\alpha = 0$, the operator A is of the form $aI + K$, where K is either a Cauchy-singular integral operator or K is compact and $a \neq 0$. The latter defines a Fredholm integral equation of the second kind, while the former defines a CSIE. In case $2\alpha > 0$, $A = L + K$, where L is a differential operator and K a possibly hypersingular integral operator. If the order of L is equal to $2\alpha > 0$, then A defines an integrodifferential equation. Otherwise, if the order of L is less than $2\alpha > 0$, we have a so-called hypersingular integral equation of the first kind. In the elasticity, (33), (38), (43), and (44) are all CSIEs. The equation (59) is a genuine Fredholm integral equation of the second kind for smooth boundary Γ with a compact operator K . Table 1 gives a quick view of the classification of BIEs based on the mapping properties of A as well as applications of BIEs in various fields.

Weak formulations for the BIEs are generally different for the first- and second-kind equations. In the former, the boundary sesquilinear forms are connected with domain sesquilinear forms for the PDEs in the interior as well as in the exterior domain, while in the latter, it connects only with the sesquilinear form either for the interior or for the exterior domain, but not both, depending on the direct or indirect approach. Also, as we have seen in the model

Table 1. Classifications and applications of BIEs.

| Classifications | Applications |
|--|--|
| Fredholm IE of the second kind $2\alpha = 0$ (e.g. 59) | Potential flow problems, viscous flows, acoustics, Darcy flows, electromagnetic fields, ... |
| CSIE $2\alpha = 0$ (e.g. 33 and 37) | Elasticity and thermoelasticity, geodesy, subsonic compressible flows, ... |
| Fredholm IE of the first kind $2\alpha < 0$ (e.g. 32 with $2\alpha = -1$) | Conformal mappings, viscous flows, acoustics, electromagnetic fields, elasticity and thermoelasticity, ... |
| Hypersingular IE $2\alpha > 0$ (e.g. 38 with $2\alpha = 1$) | Acoustics, elasticity and thermoelasticity, wings, coupling of BEM and FEM, crack problems, ... |

problem, for the second-kind BIEs, a *premultiplied operator*, as in Gatica and Hsiao (1994), is needed in order to give the appropriate duality pairing in the variational formulations for the BIEs (see (69)). As we have seen from the model problems, for the BIE (145), whose sesquilinear form coincides with the variational sesquilinear form of the BVP, the *strong ellipticity* of the boundary integral operators in the form of Gårding inequalities for the corresponding boundary integral operators in the trace space on the boundary manifold will be a consequence of *strong ellipticity* of the original BVPs (see Costabel and Wendland, 1986).

5.2 Consistency, stability, and convergence

Consistency, stability, and convergence are the basic concepts in any numerical approximating scheme. The well-known general principle, known as the Lax equivalence theorem, states that

$$\text{consistency} + \text{stability} \implies \text{convergence}$$

which applies to BEMs without any exception. In fact, Céa's lemma for the Galerkin-BEM is indeed a classical convergence theorem based on the complementary concepts of consistency and stability. In the following, let us examine the Galerkin method for the boundary integral equation (145).

Let $\mathcal{H} = H^s(\Gamma)$ denote the solution space of (145), and $\mathcal{H}_h \subset \mathcal{H}$ be a one-parameter family of finite-dimensional subspaces of \mathcal{H} . For convenience, we formulate the Galerkin

method for solving (145) in the following form: For $\sigma \in \mathcal{H}$, find $\sigma_h \in \mathcal{H}_h$ such that

$$a_T(\sigma_h, \chi_h) := (A\sigma_h, \chi_h)_{L^2(\Gamma)} = (A\sigma, \chi_h)_{L^2(\Gamma)} \quad \text{for all } \chi_h \in \mathcal{H}_h \quad (146)$$

The formulation (146) is of course equivalent to the standard one if $\sigma \in \mathcal{H}$ is the exact solution of (145). For the time being, we assume that the following conditions hold:

- (1) Consistency: Let $A_h: \mathcal{H} \rightarrow \mathcal{H}'$ be a family of continuous mappings approximating the operator A . The operators A_h are said to be consistent with A if for every $v \in \mathcal{H}$ there holds

$$\lim_{h \rightarrow 0} \|A_h v - A v\|_{\mathcal{H}'} = 0 \text{ as } h \rightarrow 0^+$$

- (2) A priori bound: For all $0 < h \leq h_0$, there exists a constant $c_0 = c_0(h_0)$ independent of σ such that

$$\|\sigma_h\|_{\mathcal{H}} \leq c_0 \|\sigma\|_{\mathcal{H}}$$

For our Galerkin method (146), we have $\mathbb{P}_h^* A_h = \mathbb{P}_h^* A \mathbb{P}_h$. Hence, consistency condition (1) is a consequence of the approximation property (110) of the sequence, while (2) is a stability condition for the family of approximate solutions. From condition (2), we see that if $\sigma = 0$, then $\sigma_h = 0$. This means that the homogeneous equation

$$(A\sigma_h, \chi_h)_{L^2(\Gamma)} = 0 \quad \text{for all } \chi_h \in \mathcal{H}_h \quad (147)$$

has only the trivial solution. Since (147) is equivalent to a quadratic system of linear equations in terms of a basis of \mathcal{H}_h , this implies the unique solvability of the inhomogeneous equation (146) for every h with $0 < h \leq h_0$. Condition (2) also implies that there is a mapping

$$\mathcal{G}_h: \mathcal{H} \ni \sigma \mapsto \sigma_h \in \mathcal{H}_h \subset \mathcal{H}$$

such that \mathcal{G}_h is uniformly bounded, that is,

$$\|\mathcal{G}_h\| \leq c_0 \quad (148)$$

Moreover, we see that $\mathcal{G}_h^* \sigma = \mathcal{G}_h \sigma_h = \sigma_h = \mathcal{G}_h \sigma$, the second equality following from the unique solvability of (146). Hence, \mathcal{G}_h is the Galerkin projection introduced in Section 4.

Now, from (148), we see that

$$A_h^{-1} := \mathcal{G}_h A^{-1}$$

is uniformly bounded, provided A^{-1} is bounded. Consequently,

$$\begin{aligned} \|\sigma - \sigma_h\|_{\mathcal{H}} &\leq c \|A_h \sigma - A_h \sigma_h\|_{\mathcal{H}'} \\ &= c \|A_h \sigma - A \sigma\|_{\mathcal{H}'} \rightarrow 0 \quad \text{as } h \rightarrow 0 \end{aligned} \quad (149)$$

as expected under Condition (1). Hence, as usual, the stability condition (2) plays a fundamental role in the abstract error estimates. We will show that stability condition (2) for the Galerkin-BEM (146) can be replaced by the well-known *Ladyženskaya-Babuška-Brezzi condition* (BBL-condition), also called *inf-sup condition*, a condition that plays a fundamental role in the study of elliptic BVPs with constraints as well as in the analysis of convergence and stability of FEMs and is most familiar to the researchers in the FEM analysis; see Nečas (1962) and Babuška and Aziz (1977).

We recall that a sesquilinear form $B(\cdot, \cdot): \mathcal{H}_1 \times \mathcal{H}_2 \rightarrow \mathbb{C}$ on Hilbert spaces \mathcal{H}_1 and \mathcal{H}_2 is said to satisfy the BBL-condition or *inf-sup condition* if there exists a constant $\gamma_0 > 0$ such that

$$\inf_{\substack{0 \neq u \in \mathcal{H}_1 \\ \|u\|_{\mathcal{H}_1} = 1}} \sup_{0 \neq v \in \mathcal{H}_2} \frac{|B(u, v)|}{\|u\|_{\mathcal{H}_1} \|v\|_{\mathcal{H}_2}} \geq \gamma_0$$

For our purpose, we consider the special discrete form of the BBL-condition with both \mathcal{H}_1 and \mathcal{H}_2 replaced by \mathcal{H} and the sesquilinear $B(\cdot, \cdot)$ form by the boundary sesquilinear form $a_T(\cdot, \cdot)$.

Definition 10 (The BBL-condition) There exists a constant $\gamma_0 > 0$ such that

$$\sup_{0 \neq \chi_h \in \mathcal{H}_h} \frac{|a_T(v_h, \chi_h)|}{\|\chi_h\|_{\mathcal{H}}} \geq \gamma_0 \|v_h\|_{\mathcal{H}} \quad \text{for all } v_h \in \mathcal{H}_h \quad (150)$$

Theorem 13. If the BBL-condition holds, then the Galerkin equations (146) are uniquely solvable for each $\sigma \in \mathcal{H}$, and we have the quasioptimal error estimate

$$\|\sigma - \sigma_h\|_{\mathcal{H}} \leq c \inf_{\chi_h \in \mathcal{H}_h} \|\sigma - \chi_h\|_{\mathcal{H}}$$

where the constant c is independent of σ and h .

Proof. If $a_T(\sigma_h, \chi_h) = 0$, then the BBL-condition (150) with $v_h = \sigma_h$ yields $\sigma_h = 0$. Then uniqueness implies the unique solvability of (146) for every $h > 0$, since (146) is equivalent to a quadratic system of linear equations. Consequently, there the mapping $\sigma \mapsto \sigma_h = \mathcal{G}_h \sigma$ is well defined, and furthermore, \mathcal{G}_h is a projection from \mathcal{H} onto $\mathcal{H}_h \subset \mathcal{H}$ by the same argument as before. It remains now to show that \mathcal{G}_h is uniformly bounded. This proceeds as

follows: For any $\sigma \in \mathcal{H}$, let σ_h correspond to its exact Galerkin solution of (146). Then we see that

$$\begin{aligned} \|\mathcal{G}_h \sigma\|_{\mathcal{H}} &= \|\sigma_h\|_{\mathcal{H}} \leq \frac{1}{\gamma_0} \sup_{0 \neq \chi_h \in \mathcal{H}_h} \frac{|a_T(\sigma_h, \chi_h)|}{\|\chi_h\|_{\mathcal{H}_h}} \\ &= \frac{1}{\gamma_0} |a_T(\sigma_h, \chi_h^*)| \\ &= \frac{1}{\gamma_0} |a_T(\sigma, \chi_h^*)| \leq c \|\sigma\|_{\mathcal{H}} \end{aligned}$$

by using the continuity of A . Here $\chi_h^* \in \mathcal{H}_h$ with $\|\chi_h^*\|_{\mathcal{H}} = 1$ denotes where the supremum on the finite-dimensional unit sphere is attained. The latter implies

$$\|\mathcal{G}_h\| := \sup_{\sigma \neq 0 \in \mathcal{H}} \frac{\|\mathcal{G}_h \sigma\|_{\mathcal{H}}}{\|\sigma\|_{\mathcal{H}}} \leq c$$

that is, (148), which implies (149) (see Chapter 9, this Volume). \square

5.3 The BBL-condition and Gårding's inequality

We have seen that the BBL-condition also plays an important role in the analysis of convergence and stability for the BEM. In the following, we would like to show that

$$\begin{aligned} &\text{Gårding's inequality} + \text{uniqueness} + (\text{ap}) \\ &\implies \text{BBL-condition} \end{aligned}$$

We need the definition of the Gårding inequality for the boundary integral operator A of (145) in the form of (88).

Definition 11 (The Gårding inequality) The boundary integral operator A is said to satisfy a Gårding inequality, if there exists a compact operator $C: \mathcal{H} \rightarrow \mathcal{H}'$ and positive constant γ such that the inequality

$$\operatorname{Re} \{a_T(v, v) + (Cv, v)_{\Gamma}\} \geq \gamma \|v\|_{\mathcal{H}}^2 \quad (151)$$

holds for all $v \in \mathcal{H}$, where \mathcal{H}' denotes the dual of \mathcal{H} .

Theorem 14. Suppose that the boundary sesquilinear form $a_T(\cdot, \cdot)$ satisfies the Gårding's inequality and

$$\begin{aligned} \operatorname{Ker}(a_T) &:= \{\sigma_0 \in \mathcal{H} | a_T(\sigma_0, \chi) = 0 \\ &\quad \text{for all } \chi \in \mathcal{H}\} = \{0\} \end{aligned}$$

Then, a_T satisfies the BBL-condition, provided \mathcal{H}_h satisfies the approximation property (ap).

Proof. Our proof follows those in Wendland (1987) (see also Wendland, 1990). From the definition of the Gårding

inequality (151), if we let $B := A + C$, then B will be \mathcal{H} -elliptic. We consider two cases:

- (1) $C = 0$ in (151). Then A is \mathcal{H} -elliptic. If we let $\chi_h = v_h$ then

$$\begin{aligned} |a_T(v_h, \chi_h)| &= |a_T(v_h, v_h)| \geq \operatorname{Re} a_T(v_h, v_h) \geq \gamma \|v_h\|_{\mathcal{H}}^2 \\ &= \gamma \|v_h\|_{\mathcal{H}} \|\chi_h\|_{\mathcal{H}} \end{aligned}$$

This implies the BBL-condition

$$\sup_{0 \neq \chi_h \in \mathcal{H}_h} \frac{|a_T(v_h, \chi_h)|}{\|\chi_h\|_{\mathcal{H}}} \geq \gamma \|v_h\|_{\mathcal{H}}$$

- (2) $C \neq 0$ in (151). Let G_{AB} be the Galerkin projection corresponding to the operator B , which is \mathcal{H} -elliptic. Then, $G_{AB} \rightarrow I$ elementwise follows from the approximation property (ap). Now let

$$L_h := I - G_{AB} B^{-1} C \quad \text{and} \quad L := I - B^{-1} C = B^{-1} A$$

Then we have

$$L - L_h = (G_{AB} - I) B^{-1} C$$

and

$$\|L - L_h\| \rightarrow 0 \quad \text{for } h \rightarrow 0$$

since B^{-1} exists and is bounded, $B^{-1} C$ is compact.

Since $L^{-1} = A^{-1} B$, the latter implies that L_h^{-1} exists and, moreover, there exists a constant c_0 independent of h for $0 < h \leq h_0$ such that

$$\|L_h^{-1}\| \leq c_0$$

that is, L_h^{-1} is uniformly bounded. Now a simple manipulation yields

$$\begin{aligned} \operatorname{Re} a_T(v_h, \chi_h) &= \operatorname{Re} (A v_h, \chi_h) = \operatorname{Re} (B L_h v_h, \chi_h) \\ &\quad - \operatorname{Re} (B(L - L_h) v_h, \chi_h) \end{aligned}$$

from which we obtain

$$|a_T(v_h, \chi_h)| + |(B(L - L_h) v_h, \chi_h)| \geq \operatorname{Re} (B L_h v_h, \chi_h) \quad (152)$$

Since B is \mathcal{H} -elliptic, if we put $\chi_h = L_h v_h$, the right-hand side of (152) is bounded as shown below:

$$\begin{aligned} \operatorname{Re} (B L_h v_h, \chi_h) &\geq \gamma_0 \|L_h v_h\|_{\mathcal{H}}^2 = \gamma_0 \|L_h v_h\|_{\mathcal{H}} \|\chi_h\|_{\mathcal{H}} \\ &\geq \frac{\gamma_0}{c_0} \|v_h\|_{\mathcal{H}} \|\chi_h\|_{\mathcal{H}} \end{aligned}$$

the last step following from the uniform boundedness of L_h^{-1} . The second term on the left-hand side of (152) is

dominated by

$$|(B(L - L_h)v_h, \chi_h)| \leq (\|B\| \|L - L_h\|) \|v_h\|_{\mathcal{H}} \|\chi_h\|_{\mathcal{H}}$$

Collecting the terms and substituting into (152), we arrive at the estimate

$$|a_F(v_h, \chi_h)| \geq \left\{ \frac{\gamma_0}{c_0} - \|B\| \|L - L_h\| \right\} \|v_h\|_{\mathcal{H}} \|\chi_h\|_{\mathcal{H}} \quad (153)$$

Since $\|L - L_h\| \rightarrow 0$, for $h \rightarrow 0$, there is a constant $h_0 > 0$ and $\gamma > 0$ independent of h for $0 < h \leq h_0$ such that

$$\left\{ \frac{\gamma_0}{c_0} - \|B\| \|L - L_h\| \right\} \geq \gamma > 0$$

The BBL-condition then follows immediately from (153). This completes the proof. \square

If the BEM is used for approximating the Dirichlet to Neumann map (or its pseudoinverse), then one needs to approximate all the Cauchy data. Then one also needs BBL-conditions between these approximating spaces in some cases. For details, see the book by Steinbach (2002).

5.4 Asymptotic error estimates

In this section, we collect some general results concerning the error estimates for the approximate solutions of (145) by the Galerkin-BEM obtained by the authors over the years. These results contain those presented in the previous sections for the model problems as special cases. We consider the boundary integral operator A of order 2α , as in (145), and assume that the following assumptions hold:

- (1) The boundary integral operator

$$A: H^{t+\alpha}(\Gamma) \mapsto H^{t-\alpha}(\Gamma)$$

defines a continuous isomorphism for any $s \in \mathbb{R}$ with $|s| \leq s_0 > 0$, that is, A is bijective and both A and A^{-1} are continuous.

- (2) The operator A satisfies a Gårding inequality of the form (151) with $\mathcal{H} = H^s(\Gamma)$ being the energy space for the operator A .
- (3) Let $\mathcal{H}_h = S_h^{s,m} \subset \mathcal{H}$ with $\ell, m \in \mathbb{N}_0$ and $m \leq \ell - 1$, that is, a regular boundary element space with approximation property and inverse property given by (132) and (133) respectively.

The following results have been established in Hsiao and Wendland (1977, 1985).

Theorem 15. Under the above assumptions, let $m > \alpha - 1/2$ for $n = 2$ or $m \geq \alpha$ for $n = 3$ and $s_0 \geq \max\{\ell, 2\alpha - \ell\}$. Then we have the asymptotic error estimate of optimal order

$$\|\sigma - \sigma_h\|_{H^t(\Gamma)} \leq ch^{t-\ell} \|\sigma\|_{H^t(\Gamma)} \quad (154)$$

for $2\alpha - \ell \leq t \leq s$, $t \leq m + 1/2$ for $n = 2$ or $t \leq m$ for $n = 3$, and $\alpha \leq s$. Moreover, the condition number of the Galerkin equations (146) is of order $O(h^{-2\alpha})$.

We note that to establish the estimates in (154) for $2\alpha - \ell \leq t \leq s \leq \ell$, we only need the approximation property of \mathcal{H}_h together with the duality arguments. On the other hand, for $\alpha \leq t \leq s \leq \ell$, we need, in addition to the approximation property, the inverse property of \mathcal{H}_h also.

For regular boundary element subspaces, (154) indicates that the rate of convergence is given by the exponent $s - t$, which is restricted by two fixed indices ℓ and m . The former indicates the degree of the complete polynomials chosen for the boundary element basis functions, while the latter governs the Sobolev index for the regularity of the solution σ . Hence, for smooth boundary, if we have a smooth datum $f \in H^{t-2\alpha}(\Gamma)$, then we have $\sigma \in H^t(\Gamma)$ from the regularity theory of pseudodifferential operators. This means that if the solution σ is sufficiently regular, we can always increase the rate of convergence by increasing the degree of the polynomials in the boundary element basis functions. On the other hand, for nonsmooth boundary, the regularity of the solution is restricted, even for given smooth data. In this case, the rate of convergence is completely unaffected, no matter how large the degree of polynomials used in the boundary element basis. Therefore, one needs more general-function spaces together with graded meshes that are not considered here.

Theorem 16. Under the same conditions as in Theorem 15, if the datum f is replaced by its L^2 -perturbation f_ϵ , then for $\alpha < 0$, we have the modified error estimate

$$\|\sigma - \sigma_h\|_{H^t(\Gamma)} \leq c[h^{t-\ell} \|\sigma\|_{H^t(\Gamma)} + h^{-(t+2\alpha)} \|f - f_\epsilon\|_{H^0(\Gamma)}]$$

If $\|f - f_\epsilon\|_{H^0(\Gamma)} \leq \epsilon$, then the choice of h given by

$$h_{\text{opt}} = \epsilon^\mu \quad \text{with } \mu := \frac{1}{s + |2\alpha|}$$

yields the optimal rate of convergence:

$$\|\sigma - \sigma_h\|_{H^t(\Gamma)} = O(\epsilon^{(s+\alpha)/(s+2\alpha)}) \quad \text{as } \epsilon \rightarrow 0^+ \quad (155)$$

We remark that the results in (155) are in agreement with those obtained by the Tikhonov-regularization technique; see Tikhonov and Arsenin (1977) and Natterer (1986).

Finally, we note that as long as the order of the boundary operator is not zero, the condition numbers of the discrete equations are unbounded, independent of the sign of α . Hence, in order to use iteration schemes for solving the discrete equations, a suitable preconditioner must be employed. On the other hand, for operators of negative orders, although the equations are ill posed, the convergence of the approximate solutions can still be achieved in view of (155).

6 CONCLUDING REMARKS

This chapter gives an overview of the Galerkin-BEM procedure as indicated in Figure 1 by using elementary model problems. Because of the limitation of the chapter length, we have confined ourselves to smooth closed boundary manifolds and omitted numerical experiments as originally planned. However, typical numerical experiments are available in Hsiao, Kopp and Wendland (1980, 1984) for illustrating the efficiency of the scheme. For nonsmooth boundary and open surfaces, see, for example, Costabel and Stephan (1985), Stephan (1987), Costabel (1988), Stephan and Wendland (1990), Hsiao, Stephan and Wendland (1991), Hsiao, Schnack and Wendland (2000), and Steinbach (2002). General error estimates for the collocation-BEM (for $n = 2$) can be found in the fundamental paper by Arnold and Wendland (1983); see also the books by Prössdorf and Silbermann (1991) and by Saranen and Vainikko (2002). For applications, see, for example, Wendland (1997).

Most of the material presented in this chapter is based on general results in a monograph, which is presently being prepared by the authors; see Hsiao and Wendland (2005) and an earlier survey by Wendland (1987). To conclude the chapter, we remark that BIEs can also be understood to be standard pseudodifferential operators on the compact boundary manifold Γ (see Seeley, 1969). Then the Gårding inequality is equivalent to the positive definiteness of the principal symbol of pseudodifferential operators. For particularly chosen BIEs such as the ones for our model problems, the Gårding inequality can be obtained from the coerciveness of the original strongly elliptic BVP. For the convergence of the Galerkin and the collocation BEMs, we only require the principal symbol to be positive definite modulo multiplications by regular functions or matrices. This is the definition of strong ellipticity introduced in Stephan and Wendland (1976) for the systems

of pseudodifferential equations. Details of these concepts and relations will be available in the monograph Hsiao and Wendland (2005), which also contains rather complete theoretical results for a class of boundary integral operators having symbols of rational type (see also McLean, 2000). The latter cover almost all boundary integral operators in applications.

ACKNOWLEDGMENTS

This work was supported by the DFG priority research programme 'Boundary Element Methods' within the guest programme We-659/19-2, and by the 'Graduiertenkolleg Modellierung und Diskretisierungsmethoden für Kontinua und Störungen' at the Universität Stuttgart. The work of GCH was also partially supported by the Weierstrass-Institut für Angewandte Analysis und Stochastik im Forschungsverbund Berlin e.V. The work of WLW was also partially supported when he was a Visiting Research Scholar at the University of Delaware in 2002.

REFERENCES

- Adams DA. *Sobolev Spaces*. Academic Press: New York, 1975.
- Anselone PM. *Collectively Compact Operator Approximation Theory and Applications to Integral Equations*. Prentice Hall: Englewood Cliffs, 1971.
- Arnold DN and Wendland WL. On the asymptotic convergence of collocation methods. *Math. Comp.* 1983; 41:197–242.
- Aubin JP. *Approximation of Elliptic Boundary Value Problems*. Wiley-Interscience: New York, 1972.
- Babuška I and Aziz AK. *Integral Equation Methods in Potential Theory and Elastostatics*. Academic Press: New York, 1977.
- Bers L, John F and Schechter M. *Partial Differential Equations*. John Wiley & Sons: New York, 1964.
- Clairier PG. *The Finite Element Method for Elliptic Problems*. North Holland: Amsterdam, 1978.
- Costabel M. Boundary integral operators on Lipschitz domains: elementary results. *SIAM J. Math. Anal.* 1988; 19: 613–626.
- Costabel M and Stephan EP. Boundary integral equations for mixed boundary value problems in polygonal domains and Galerkin approximation. *Mathematical Models in Mechanics*, vol. 15. Banach Center Publications: Warsaw, 1985; 175–251.
- Costabel M and Stephan EP. Duality estimates for the numerical approximation of boundary integral equations. *Numer. Math.* 1988; 54: 339–353.
- Costabel M and Wendland WL. Strong ellipticity of boundary integral operators. *J. Reine Angew. Math.* 1986; 372: 34–63.
- Dautray R and Lions JL. *Mathematical Analysis and Numerical Methods for Science and Technology*, vol. 4. Springer-Verlag: Berlin, 1990.

- Fichera G. Linear elliptic equations of higher order in two independent variables and singular integral equations, with applications to anisotropic inhomogeneous elasticity. In *Partial Differential Equations and Continuum Mechanics*, Langer RE (ed.), The University of Wisconsin Press: Wisconsin, 1961; 55–80.
- Fischer TM, Hsiao GC and Wendland WL. Singular perturbations for the exterior three-dimensional slow viscous flow problem. *J. Math. Anal. Appl.* 1985; 110: 583–603.
- Gatica GN and Hsiao GC. A Gårding's inequality for variational problems with constraints. *Appl. Anal.* 1994; 54: 73–90.
- Hess JL and Smith AMO. Calculation of potential flow about arbitrary bodies. In *Progress in Aeronautical Sciences*, Kuchemann D (ed.), vol. 8. Pergamon Press: Oxford, 1966; 1–138.
- Hildebrandt S and Wienholtz E. Constructive proofs of representation theorems in separable Hilbert space. *Commun. Pure Appl. Math.* 1964; 17: 369–373.
- Hsiao GC. On the stability of integral equations of the first kind with logarithmic kernels. *Arch. Ration. Mech. Anal.* 1986; 94: 179–192.
- Hsiao GC. On the stability of boundary element methods for integral equations of the first kind. In *Boundary Elements IX*, Brebbia CA, Wendland WL and Kuhn G (eds), Springer-Verlag: New York, 1987; 177–191.
- Hsiao GC. Variational methods for boundary integral equations: theory and applications. In *Problemi Attuali Dell'Analisi e Della Fisica Matematica*, Ricci PE (ed.), Aracne Editrice: Roma, 2000; 59–76.
- Hsiao G and MacCamy RC. Solution of boundary value problems by integral equations of the first kind. *SIAM Rev.* 1973; 15: 687–705.
- Hsiao GC and Wendland WL. A finite element method for some integral equations of the first kind. *J. Math. Anal. Appl.* 1977; 58: 449–481.
- Hsiao GC and Wendland WL. The Aubin–Nitsche lemma for integral equations. *J. Integral Equations* 1981a; 3: 299–315.
- Hsiao GC and Wendland WL. Super-approximation for boundary integral methods. In *Advances in Computer Methods for Partial Differential Equations-IV*, Vichnevetsky R and Stepleman RS (eds), IMACS: New Brunswick N.J., 1981b; 200–205.
- Hsiao GC and Wendland WL. On a boundary integral method for some exterior problems in elasticity. *Math. Mech. Astron.* 1985; 257(V): 31–60.
- Hsiao GC and Wendland WL. *Boundary Integral Equations: Variational Methods*. Springer-Verlag: Heidelberg, 2005.
- Hsiao GC, Kopp P and Wendland WL. Galerkin collocation method for some integral equations of the first kind. *Computing* 1980; 25: 299–315.
- Hsiao GC, Kopp P and Wendland WL. Some applications of a Galerkin-collocation method for boundary integral equations of the first kind. *Math. Methods Appl. Sci.* 1984; 6: 280–325.
- Hsiao GC, Schnack E and Wendland WL. Hybrid finite-boundary element methods for elliptic systems of second order. *Comput. Methods Appl. Mech. Eng.* 2000; 190: 431–485.
- Hsiao GC, Stephan EP and Wendland WL. On the Dirichlet problem in elasticity for a domain exterior to an arc. *J. Comput. Appl. Math.* 1991; 34: 1–19.
- Jaswon MA and Symm GT. *Integral Equation Methods in Potential Theory and Elastostatics*. Academic Press: London, 1977.
- Kantorovich LY and Akilov GP. *Functional Analysis in Normed Spaces*. Pergamon Press: New York, 1964 (Russian edition: Fizmatgiz: Moscow, 1959).
- Kuhn M and Steinbach O. Symmetric coupling of finite and boundary elements for exterior magnetic field problems. *Math. Methods Appl. Sci.* 2002; 25: 357–371.
- Kupradze VD. *Potential Methods in the Theory of Elasticity*. Israel Program Scientific Transl.: Jerusalem, 1965.
- Kupradze VD, Gegelia TG, Bacheleishvili MO and Burchuladze TV. *Three-Dimensional Problems of the Mathematical Theory of Elasticity and Thermoelasticity*. North Holland: Amsterdam, 1979.
- Lax PD and Milgram AN. Parabolic equations, contribution to the theory of partial differential equations. *Ann. Math.* 1954; 33: 167–190.
- LeRoux MN. Méthode d'élément finis pour la résolution numérique de problèmes extérieurs en dimension 2. *RAIRO Anal. Numer.* 1977; 11: 27–60.
- Lions JL and Magenes E. *Nonhomogeneous Boundary Value Problems and Applications*, Vol. I–III. Springer-Verlag: Berlin, 1972.
- MacCamy RC. On a class of two-dimensional Stokes flows. *Arch. Ration. Mech. Anal.* 1966; 21: 256–258.
- McLean W. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press: Cambridge, 2000.
- Meyers N and Serrin J. H = W. *Proc. Natl. Acad. Sci. U.S.A.* 1964; 51: 1055–1056.
- Mikhlin SG. *Multidimensional Singular Integrals and Integral Equations*. Pergamon Press: New York, 1965 (Russian edition: Fizmatgiz: Moscow, 1962).
- Mikhlin SG. *Variationsmethoden der Mathematischen Physik*. Akademie-Verlag: Berlin, 1970.
- Mikhlin SG and Prössdorf S. *Singular Integral Operators*. Springer-Verlag: Berlin, 1986.
- Millman RS and Parker GD. *Elements of Differential Geometry*. Prentice Hall: Englewood Cliffs, 1977.
- Muskhlishvili NI. *Singular Integral Equations*. Noordhoff: Groningen, 1953.
- Natterer F. *The Mathematics of Computerized Tomography*. John Wiley & Sons: Chichester, 1986.
- Nečas J. Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle. *Ann. Scuola Norm. Sup. Pisa* 1962; 16: 305–326.
- Nečas J. *Les Méthodes Directes en Théorie des Équations Elliptiques*. Masson: Paris, 1967.
- Nedelec JC. Curved finite element methods for the solution of singular integral equations on surfaces in R^3 . *Comput. Methods Appl. Mech. Eng.* 1976; 9: 191–216.
- Nedelec JC. Approximation des équations intégrales en mécanique et en physique. *Lecture Notes*. Centre de Mathématiques Appliquées, École Polytechnique, 91128 Palaiseau Cedex, France, 1977.
- Nedelec JC and Planchard J. Une méthode variationnelle d'éléments finis pour la résolution numérique d'un problème extérieur dans R^3 . *RAIRO Anal. Numer.* 1973; 7: 105–129.
- Nitsche JA. Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens. *Numer. Math.* 1968; 11: 346–348.
- Prössdorf S and Silbermann B. *Numerical Analysis for Integral and Related Operator Equations*. Birkhäuser-Verlag: Basel, 1991.
- Saranen J and Vainikko G. *Periodic Integral and Pseudodifferential Equations with Numerical Application*. Springer-Verlag: Berlin, 2002.
- Seeley RT. Topics in pseudodifferential operators. In *Pseudodifferential Operators*, Nirenberg L (ed.). CIME Cremonese: Roma, 1969; 169–305.
- Steinbach O. On the stability of the L2-projection in fractional Sobolev spaces. *Numer. Math.* 2001; 88: 367–379.
- Steinbach O. *Stability Estimates for Hybrid Coupled Domain Decomposition Methods*. Springer-Verlag: Heidelberg, 2002.
- Steinbach O and Wendland WL. The construction of some efficient preconditioners in the boundary element method. *Adv. Comput. Math.* 1998; 9: 191–216.
- Steinbach O and Wendland WL. On C. Neumann's method for second-order elliptic systems in domains with non-smooth boundaries. *J. Math. Anal. Appl.* 2001; 262: 733–748.
- Stephan EP. Boundary integral equations for screen problems in R^3 . *Integral Equations Operator Theory* 1987; 10: 236–257.
- Stephan E and Wendland WL. Remarks to Galka and least squares methods with finite elements for general elliptic problems. *Manuscr. Geodactica* 1976; 1: 93–123.
- Stephan E and Wendland WL. A hypersingular boundary integral method for two-dimensional screen and crack problems. *Archive Ration. Mech. Anal.* 1990; 112: 363–390.
- Taylor ME. *Pseudodifferential Operators*. Princeton University Press: Princeton, 1981.
- Tikhonov AN and Arsenin VY. *Solutions of Ill-Posed Problems*. John Wiley & Sons: Chichester, 1977.
- Wendland WL. Lösung der ersten und zweiten Randwertaufgaben des Innen- und Aussengebietes für die Potentialgleichung im R^3 durch Randbelegungen. Doctoral Thesis, TU: Berlin, D83; 1965.
- Wendland WL. Boundary element methods and their asymptotic convergence. In *Theoretical Acoustics and Numerical Techniques*, Filippi P (ed.). CISM No. 277. Springer-Verlag: Wien, 1983; 155–216.
- Wendland WL. Strongly elliptic boundary integral equations. In *The State of the Art in Numerical Analysis*, Iscris A and Powell MID (eds). Clarendon Press: Oxford, 1987; 511–562.
- Wendland WL. Boundary element methods for elliptic problems. In *Mathematical Theory of Finite and Boundary Element Methods*, Schatz AH, Thomé V and Wendland WL (eds). Birkhäuser-Verlag: Basel, 1990; 219–276.
- Wendland WL (ed.). *Boundary Element Topics*. Springer-Verlag: Berlin, 1997.
- Atkinson K and Han W. *Theoretical Numerical Analysis: A Functional Analysis Framework*. Springer-Verlag: New York, 2001.
- Banerjee PK. *The Boundary Element Methods in Engineering*. McGraw-Hill: New York, 1994.
- Brebbia CA and Dominguez J. *Boundary Elements: An Introductory Course*. McGraw-Hill: New York, 1992.
- Chen G and Zhou J. *Boundary Element Methods*. Academic Press: London, 1992.
- Elschner J. *Singular Ordinary Differential Operators and Pseudodifferential Equations*. Akademie-Verlag: Berlin, 1985.
- Gatica N and Hsiao GC. *Boundary-field Equation Methods for a Class of Nonlinear Problems*. Pitman Research Notes in Mathematics Series 331. Addison-Wesley Longman: Edinburgh Gate, Harlow, 1995.
- Hartmann F. *Introduction to Boundary Elements*. Springer-Verlag: Berlin, 1989.
- Kinderlehrer D and Stampacchia G. *An Introduction to Variational Inequalities and their Application*. Academic Press: London, 1980.
- Kress R. *Linear Integral Equations*. Springer-Verlag: New York, 1989.
- Martensen E. *Potentialtheorie*. B.G. Teubner: Stuttgart, 1968.
- Maz'ya VG. Boundary integral equations. In *Encyclopaedia of Mathematical Sciences*, vol. 27, Analysis IV. Maz'ya VG and Nikolskii SM (eds). Springer-Verlag: Berlin, 1991; 127–222.
- Mikhlin SG. *Mathematical Physics, An Advanced Course*. North Holland: Amsterdam, 1970.
- Mikhlin SG. *Partielle Differentialgleichungen in der Mathematischen Physik*. Akademie-Verlag: Berlin, 1978.
- Nedelec JC. *Acoustic and Electromagnetic Equations*. Springer-Verlag: New York, 2001.
- Schmeidler W. *Integralgleichungen mit Anwendungen in Physik und Technik*. Akademische Verlagsgesellschaft Geest & Poritz: Leipzig, 1950.
- Stein E and Wendland WL (eds). *Finite Element and Boundary Element Techniques from Mathematical and Engineering Point of View*. CISM No. 501. Springer-Verlag: Wien, 1988.
- Verfürth R. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. John Wiley & Sons/B.G. Teubner: Chichester/Stuttgart, 1996.
- Wendland WL. Die Behandlung von Randwertaufgaben in R^3 mit Hilfe von Einfach- und Doppelschichtpotentialen. *Numer. Math.* 1968; 11: 380–404.
- Wendland WL. *Elliptic Systems in the Plane*. Pitman: London, 1979.
- Yu De-hao. *Natural Boundary Integral Method and its Applications*. Science Press/Kluwer Academic Publishers: Dordrecht, 2002.
- Zhu J. *Boundary Element Analysis for Elliptic Boundary value Problems* (in Chinese). Science Press: Beijing, 1991.

Chapter 13

Coupling of Boundary Element Methods and Finite Element Methods

Ernst P. Stephan

Institut für Angewandte Mathematik, Universität Hannover, Hannover, Germany

| | |
|--|-----|
| 1 Introduction | 375 |
| 2 Symmetric Coupling of Standard Finite Elements and Boundary Elements | 377 |
| 3 Fast Solvers for the hp-version of FE/BE Coupling | 389 |
| 4 Least Squares FE/BE Coupling Method | 394 |
| 5 FE/BE Coupling for Interface Problems with Signorini Contact | 396 |
| 6 Applications | 403 |
| 7 Concluding Remarks | 408 |
| References | 409 |

1 INTRODUCTION

The coupling of boundary element methods (BEM) with finite element methods (FEM) has become a very powerful and popular method in applications, and there exist detailed descriptions of implementations of such methods (see Brebbia, Telles and Wrobel (1984) and the references given there). Within engineering computations, the BEM and the FEM are well-established tools for the numerical approximation of real-life problems in which analytical solutions are mostly unknown or available under unrealistic modeling only.

Both methods are somehow complementary: the FEM seems to be more general and applicable to essentially

nonlinear problems, while the BEM is restricted to certain linear problems with constant coefficients. On the other hand, the FEM requires a bounded domain, while the BEM models an unbounded exterior body as well. Applications occur in scattering problems, elastodynamics, electromagnetism, and elasticity. The general setting of these problems is that we wish to solve a given differential equation in two adjacent domains subject to a specified interface condition. Most often, we have a bounded region Ω surrounded by an unbounded region, with the interface condition being specified on the shared boundary. Typically, in this 'marriage à la mode' (see Bettess, Kelly and Zienkiewicz, 1977, 1979), an FEM formulation is used to describe the solution within the bounded region – where the differential equation can be nonlinear – and the BEM is used to represent the exterior solution.

The purpose of the present note is to give an overview of FE/BE coupling procedures for elliptic boundary value problems. Special emphasis is given to describe the symmetric coupling method, which was independently proposed and analyzed for linear transmission problems by Costabel and Stephan (see Costabel (1988b) for general strongly elliptic second-order systems, see Costabel (1987) for higher-order systems, and see Costabel and Stephan (1988a) and Costabel and Stephan (1988b) for the case of scattering of elastic waves) and by Han (see Han, 1990); later, this method was extended to interface problems with nonlinear differential equations in Ω (see Costabel and Stephan, 1990; Gatica and Hsiao, 1989, 1990, 1995). This symmetric coupling method (see Section 2) has been described in the engineering literature for problems from elasticity and elastoplasticity in Polizzotto (1987) and allows a variational formulation

in which the solution satisfies a saddle-point problem. In this method, the set of boundary integral operators consists of the single-layer potential, the double-layer potential, and their normal derivatives. The computed Galerkin approximations, consisting of finite elements (FE) in Ω coupled with boundary elements (BE) on the Lipschitz continuous interface boundary Γ , converge quasioptimally in the energy norm. This is an advantage over other coupling methods of a different structure; see Bielak and MacCamy (1983) and Johnson and Nedelec (1980). These coupling methods cannot guarantee convergence of the Galerkin solutions for nonlinear interface problems, for example, with Hencky-von Mises-type material considered in Section 6.1. As against these coupling methods, one uses the hypersingular operator in the symmetric method; this operator is given by the normal derivative of the double-layer potential. Although traditionally avoided if possible, this operator is now considered as one of the classical boundary integral operators and it appears frequently in the treatment of crack problems. It is known that the apparent difficulties in the numerical treatment of this operator can be easily overcome; see Maue (1949) and Nedelec (1982).

In Section 2.1, we present a mathematical analysis of the symmetric coupling method based on the strong coerciveness of the Steklov–Poincaré operator and its discrete analogue. Adaptive versions of the FE/BE coupling method are presented using error indicators of residual type and those based on hierarchical two-level subspace decompositions. Also, implicit estimators are given in which local Neumann problems in the FEM domain are solved.

Another prime topic of this article is the iterative solution of the discrete systems resulting from the above symmetric FE/BE coupling. These systems have block structured symmetric and indefinite matrices. An important approach to the construction of a fast solver has been the use of the Schur complement (see Bramble and Pasciak, 1988) in connection with domain decomposition techniques (see Langer, 1994; Carstensen, Kuhn and Langer, 1998; and Steinbach and Wendland, 1997). As the symmetric coupling method leads to a symmetric and positive definite Schur complement, it can be treated by very fast standard preconditioners. In order to avoid the direct computation of the inverse of the discretization of the single-layer potential, which is involved in the Schur complement, one might turn to nested inner–outer iterations (see Hahne, Stephan and Thies, 1995). The alternative we propose here is to use the minimum or conjugate residual method (MINRES), which is an ‘optimal’ solver belonging to the family of Krylov-subspace (conjugate-gradient-like) iterative methods. In Section 3, we consider the uniform hp-version and apply MINRES with multilevel preconditioning for the FE block and the discretized version of the single-layer

potential. We identify the spectral properties of the components of the system matrix and give estimates for the iteration numbers and comment on implementation issues. As in Heuer, Maischak and Stephan (1999), we show that the eigenvalues of the preconditioned Galerkin matrix are appropriately bounded in case of the two-block preconditioner or depend mildly on the grid size h and the polynomial degree p of the trial functions. Especially, the use of additive Schwarz preconditioners leads to efficient solvers. Other good choices as preconditioners are multigrid (MG), BPX, or those using hierarchical bases. Alternatively, by simple scaling of BE test functions, we obtain a system with nonsymmetric Galerkin matrix with positive (semi-)definite symmetric part, which can be solved by the generalized minimal residual method (GMRES).

A current research topic is the least squares coupling of FEs and BEs that allows to deal with mixed FE formulations without demanding inf–sup conditions to hold. This recent development is considered in Section 4. An increasing interest has evolved in applying mixed methods instead of usual FEM together with either boundary integral equations or Dirichlet-to-Neumann (DtN) mappings; see also Hsiao, Steinbach and Wendland (2000) and Steinbach (2003). Often in applications, a mixed FEM is more beneficial than the standard FEM, for example, in structural mechanics via mixed methods, stresses are computed more accurately than displacements. However, in such mixed FE/BE coupling methods, it is often difficult to work with finite element spaces that satisfy appropriate discrete inf–sup conditions. On the other hand, least squares formulations do not require inf–sup conditions to be satisfied. Therefore, they are especially attractive to use in combination with mixed formulations. Recently, in Bramble, Lazarov and Pasciak (1997), a least squares functional was introduced, involving a discrete inner product related to the inner product in the Sobolev space of order -1 . This approach is extended to a least squares coupling with boundary integral operators in Gatica, Harbrecht and Schneider (2001) and Maischak and Stephan. Discrete versions of the inner products in the Sobolev spaces $\tilde{H}^{-1}(\Omega)$ and $H^{1/2}(\Gamma)$ are constructed by applying multigrid (MG) or BPX to both FE and BE discretizations.

Another research topic is the FE/BE coupling for Signorini-type interface problems, which leads to variational inequalities with boundary integral operators. For the BEM applied to variational inequalities, see Hsiao and Han (1988), Gwinner and Stephan (1993), Spann (1993), and Eck *et al.* (2003). In Section 5, we propose two approaches for the FE/BE coupling for Signorini-type interface problems, namely, a primal method and

a dual-mixed method. For the h-version of the primal-coupling method, existence, uniqueness, and convergence results were obtained by Carstensen and Gwinner (1997). Maischak (2001) extends their approach to the hp-version and derives a posteriori error estimates based on hierarchical subspace decompositions for FE and BE, similar to Mund and Stephan (1999) for nonlinear transmission problems. A dual-mixed coupling method for the h-version, which heavily uses the inverse of the Steklov–Poincaré operator and its discrete versions, is analyzed in Maischak (2001) (see also Gatica, Maischak and Stephan, 2003). Both primal- and dual-mixed coupling methods are described in this article for the scalar case; the extension to corresponding elasticity problems with Signorini-type interface conditions involves no major difficulties. FE/BE coupling procedures for friction problems are currently under investigation and are not considered here. As system solvers for the discretized variational inequalities, we propose a preconditioned Polyak algorithm in the case of the primal method and a preconditioned modified Uzawa algorithm in the case of the dual-mixed method. Both solvers lead to efficient numerical procedures, even for adaptively refined meshes.

Section 6 deals with the applications of FE/BE coupling methods to elasticity problems. Firstly, we describe the symmetric FE/BE coupling method for a nonlinear Hencky–von Mises stress–strain relation. Here, we comment on the saddle-point structure of the symmetric coupling method. In the symmetric coupling method, the displacements in the interior domain are H^1 -regular and the equilibrium of tractions across the interface is satisfied weakly. In applications, however, the stresses are often more important to determine than displacements. Then, the coupling of BEs and mixed FEs, in which approximations to stresses can be determined directly, is more adequate; see Brink, Carstensen and Stein (1996) and Meddahi *et al.* (1996). Secondly, therefore, following Brink, Carstensen and Stein (1996), we present here a dual-mixed formulation for the finite element part Ω_F of a model problem in plane linear elasticity from Gatica, Heuer and Stephan (2001). This means that the stresses σ are required to satisfy $\sigma \in [L^2(\Omega_F)]^{2 \times 2}$ and $\text{div } \sigma \in [L^2(\Omega_F)]^2$, whereas the displacements are only sought in $[L^2(\Omega_F)]^2$. Thus, the approximate tractions are continuous across element sides and the interface boundary, while the displacements are continuous only in a weak sense across element sides and the interface boundary. We present an a posteriori error estimator, which is based on the solution of local elliptic problems with Dirichlet boundary conditions. We note that the a posteriori error estimate can be derived in an analogous way for three-dimensional elasticity problems. For problems with nonlinearities, however, the inversion

of the elasticity tensor C is not adequate. In that case, a different so-called dual–dual formulation of the problem can be used (see Section 7), in which further references are also given to other topics not explicitly considered here.

2 SYMMETRIC COUPLING OF STANDARD FINITE ELEMENTS AND BOUNDARY ELEMENTS

For a model interface problem, we present a combined approach with FE and BE. The given symmetric coupling method renders all boundary conditions on the interface manifold Γ to be natural and also allows for a nonlinear elliptic differential operator in the bounded domain Ω_1 . Our solution procedure makes use of an integral equation method for the exterior problem and of an energy (variational) method for the interior problem and consists of coupling both methods via the transmission conditions on the interface. We give an equivalence result for the solution satisfying a weak coupling formulation. We solve this variational formulation with the Galerkin method using FE in Ω_1 and BE on Γ . As shown in Wendland (1988) and Costabel and Stephan (1990), we have convergence and quasioptimality of the Galerkin error in the energy norm. At the end of this section, we comment on the exponential convergence rate of the hp-version of the coupling procedure when an appropriate geometric mesh refinement is used.

Let $\Omega_1 := \Omega \subset \mathbb{R}^d$, $d \geq 2$ be a bounded domain with Lipschitz boundary $\Gamma = \partial\Omega_1$, and $\Omega_2 := \mathbb{R}^d \setminus \bar{\Omega}_1$ with normal n on Γ pointing into Ω_2 . For given $f \in L^2(\Omega_1)$, $u_0 \in H^{1/2}(\Gamma)$, $t_0 \in H^{-1/2}(\Gamma)$, we consider the following model interface problem (IP):

Problem (IP): Find $u_1 \in H^1(\Omega_1)$, $u_2 \in H_{\text{loc}}^1(\Omega_2)$ such that

$$-\text{div } A(\nabla u_1) = f \quad \text{in } \Omega_1 \quad (1)$$

$$\Delta u_2 = 0 \quad \text{in } \Omega_2 \quad (2)$$

$$u_1 = u_2 + u_0 \quad \text{on } \Gamma \quad (3)$$

$$A(\nabla u_1) \cdot n = \frac{\partial u_2}{\partial n} + t_0 \quad \text{on } \Gamma \quad (4)$$

$$u_2(x) = \begin{cases} b \log |x| + o(1), & d = 2, \\ O(|x|^{2-d}), & d \geq 3, \end{cases} \quad |x| \rightarrow \infty \quad (5)$$

where $b \in \mathbb{R}$ is a constant (depending on u_2). The operator A is assumed to be uniformly monotone and Lipschitz continuous, that is, there exist positive constants α and C

such that for all $\eta, \tau \in L^2(\Omega)^d$

$$\int_{\Omega} (A(\eta) - A(\tau)) \cdot (\eta - \tau) dx \geq \alpha \|\eta - \tau\|_{0,\Omega}^2 \quad (6)$$

$$\|A(\eta) - A(\tau)\|_{0,\Omega} \leq C \|\eta - \tau\|_{0,\Omega} \quad (7)$$

Here $\|\cdot\|_{0,\Omega}$ denotes the norm in $L^2(\Omega)^d$. Examples of operators of this type can be found in Stephan (1992) and for elasticity in Zeidler (1988, Section 62).

The definition of the Sobolev spaces is as usual:

$$H^s(\Omega) = \{\phi|_{\Omega}; \phi \in H^s(\mathbb{R}^d)\} \quad (s \in \mathbb{R})$$

$$H^s(\Gamma) = \begin{cases} \{\phi|_{\Gamma}; \phi \in H^{s+1/2}(\mathbb{R}^d)\} & (s > 0) \\ L^2(\Gamma) & (s = 0) \\ (H^{-s}(\Gamma))' \text{ (dual space)} & (s < 0) \end{cases}$$

In the following, we often write $\|\cdot\|_{s,B}$ for the Sobolev norm $\|\cdot\|_{H^s(B)}$ with $B = \Omega$ or Γ .

We now derive the symmetric coupling method (see Costabel, 1988b; Costabel and Stephan, 1990) as discussed in detail in Costabel, Ervin, Stephan (1991). By using Green's formula together with the decaying condition (5), one is led to the representation formula for the solution in the exterior domain u_2 of (2)

$$u_2(x) = \int_{\Gamma} \left\{ \frac{\partial}{\partial n(y)} G(x, y) u_2(y) - G(x, y) \frac{\partial u_2}{\partial n(y)} \right\} ds_y, \quad x \in \Omega_2 \quad (8)$$

with the fundamental solution of the Laplacian given by

$$G(x, y) = \begin{cases} -\frac{1}{\omega_2} \log |x - y|, & d = 2 \\ \frac{1}{\omega_d} |x - y|^{2-d}, & d \geq 3 \end{cases} \quad (9)$$

where we have $\omega_2 = 2\pi$, $\omega_3 = 4\pi$.

By using the boundary integral operators

$$V\psi(x) := 2 \int_{\Gamma} G(x, y) \psi(y) ds_y, \quad x \in \Gamma \quad (10)$$

$$K\psi(x) := 2 \int_{\Gamma} \frac{\partial}{\partial n_y} G(x, y) \psi(y) ds_y, \quad x \in \Gamma \quad (11)$$

$$K'\psi(x) := 2 \frac{\partial}{\partial n_x} \int_{\Gamma} G(x, y) \psi(y) ds_y, \quad x \in \Gamma \quad (12)$$

$$W\psi(x) := -2 \frac{\partial}{\partial n_x} \int_{\Gamma} \frac{\partial}{\partial n_y} G(x, y) \psi(y) ds_y, \quad x \in \Gamma \quad (13)$$

together with their well-known jump conditions, we obtain from (8) the following integral equations:

$$2 \frac{\partial u_2}{\partial n} = -Wu_2 + (I - K') \frac{\partial u_2}{\partial n} \quad (14)$$

$$0 = (I - K)u_2 + V \frac{\partial u_2}{\partial n} \quad (15)$$

In order to give an equivalent formulation for (IP), we use from Costabel (1988a) the following mapping properties of the boundary integral operators in which the duality (\cdot, \cdot) between the spaces $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$ extends the scalar product in $L^2(\Gamma)$.

Lemma 1. (a) Let $\Gamma = \partial\Omega$ be a Lipschitz boundary. The operators

$$\begin{aligned} V: H^{-1/2}(\Gamma) &\longrightarrow H^{1/2}(\Gamma) \\ K: H^{1/2}(\Gamma) &\longrightarrow H^{1/2}(\Gamma) \\ K': H^{-1/2}(\Gamma) &\longrightarrow H^{-1/2}(\Gamma) \\ W: H^{1/2}(\Gamma) &\longrightarrow H^{-1/2}(\Gamma) \end{aligned}$$

are continuous. Moreover, the single-layer potential V and the hypersingular operator W are symmetric; the double-layer potential K has the dual K' .

(b) For $d = 2$ and provided the capacity of Γ , $\text{cap}(\Gamma)$, is less than 1, or $d = 3$, then $V: H^{-1/2}(\Gamma) \rightarrow H^{1/2}(\Gamma)$ is positive definite, that is, there is a constant $\alpha > 0$ such that

$$(\varphi, V\varphi) \geq \alpha \|\varphi\|_{-1/2,\Gamma}^2 \quad \forall \varphi \in H^{-1/2}(\Gamma) \quad (16)$$

(c) The kernel of the operator W consists of the constant functions. W is positive semidefinite, that is,

$$(v, Wv) \geq 0 \quad \forall v \in H^{1/2}(\Gamma) \quad (17)$$

Remark 1. For the definition of $\text{cap}(\Gamma)$, we refer to Gaier (1976) and Sioan and Spence (1988), and we only mention here that if Ω lies in a ball with radius less than 1, for example, then $\text{cap}(\Gamma) < 1$. Thus, $\text{cap}(\Gamma) < 1$ can always be achieved by scaling; see Hsiao and Wendland (1977) and Stephan and Wendland (1984).

Next, we derive the symmetric coupling method for (IP). One observes that (15) forms one part of the weak formulation of problem (IP)

$$\begin{aligned} -(u_1, \psi) - \left(V \frac{\partial u_2}{\partial n}, \psi \right) + (Ku_1, \psi) \\ = -(u_0, \psi) + (Ku_0, \psi) \quad \forall \psi \in H^{-1/2}(\Gamma) \end{aligned} \quad (18)$$

The second part of the weak formulation has to couple the exterior problem (2) and the interior problem (1). The weak formulation of the latter is

$$\begin{aligned} a(u_1, v) := \int_{\Omega_1} A(\nabla u_1) \cdot \nabla v dx = \int_{\Gamma} (A(\nabla u_1) \cdot n) v ds \\ + \int_{\Omega_1} f v dx \quad \forall v \in H^1(\Omega_1) \end{aligned} \quad (19)$$

Taking the integral equation (14) and substituting (3) and (4) into (19), one obtains

$$\begin{aligned} 2a(u_1, v) - \left(\frac{\partial u_2}{\partial n}, v \right) + \left(K' \frac{\partial u_2}{\partial n}, v \right) + (Wu_1, v) = 2(f, v) \\ + 2(t_0, v) + (Wu_0, v) \quad \forall v \in H^1(\Omega_1) \end{aligned} \quad (20)$$

where $(f, v) = \int_{\Omega_1} f v dx$. There holds the following equivalence.

Theorem 1. If u_1 and u_2 solve (1)–(5), then u_1 and $\partial u_2 / \partial n$ satisfy (18) and (20). Conversely, provided u_1 and $\partial u_2 / \partial n$ solve (18) and (20), then u_1 and u_2 , which is defined by (8), (3), and (4), are the solutions of the interface problem (IP).

Note that in this way we obtain the following variational formulation.

Find $u := u_1 \in H^1(\Omega_1)$ and $\phi := (\partial u_2 / \partial n) \in H^{-1/2}(\Gamma)$ such that for all $v \in H^1(\Omega_1)$ and $\psi \in H^{-1/2}(\Gamma)$

$$\begin{aligned} 2a(u, v) + ((K' - I)\phi, v) + (Wu, v) \\ = 2(t_0, v) + (Wu_0, v) + 2(f, v) \\ ((K - I)u, \psi) - (V\phi, \psi) = ((K - I)u_0, \psi) \end{aligned} \quad (21)$$

For the Galerkin scheme, we choose finite-dimensional subspaces $X_M \subset H^1(\Omega_1)$ and $Y_N \subset H^{-1/2}(\Gamma)$ and define the Galerkin solution $(u_M, \phi_N) \in X_M \times Y_N$ by

$$\begin{aligned} 2a(u_M, v) + ((K' - I)\phi_N, v) + (Wu_M, v) \\ = 2(t_0, v) + (Wu_0, v) + 2(f, v) \\ ((K - I)u_M, \psi) - (V\phi_N, \psi) = ((K - I)u_0, \psi) \end{aligned} \quad (22)$$

for all $v \in X_M$ and $\psi \in Y_N$.

There holds the following convergence result.

Theorem 2. Every Galerkin scheme (22) with approximating finite-dimensional spaces $X_M \subset H^1(\Omega_1)$ and $Y_N \subset H^{-1/2}(\Gamma)$ converges with optimal order, that is, with the exact solution (u, ϕ) of (21) and the Galerkin solution

(u_M, ϕ_N) of (22), there holds the estimate

$$\begin{aligned} \|u - u_M\|_{1,\Omega} + \|\phi - \phi_N\|_{-1/2,\Gamma} \\ \leq C \left\{ \inf_{\tilde{u} \in X_M} \|u - \tilde{u}\|_{1,\Omega} + \inf_{\tilde{\phi} \in Y_N} \|\phi - \tilde{\phi}\|_{-1/2,\Gamma} \right\} \end{aligned} \quad (23)$$

where the constant C is independent of M, N, u , and ϕ .

For a slightly modified interface problem with

$$-\Delta u_1 + u_1 = f \text{ in } \Omega_1 \quad (24)$$

the convergence of the Galerkin scheme of the FE/BE coupling follows, owing to Stephan and Wendland (1976), directly from the strong ellipticity of the system (21). Choosing $v = u$ and $\psi = -\phi$ shows that the inf-sup condition is satisfied (Costabel and Stephan, 1988b), namely,

$$\begin{aligned} \|u\|_{1,\Omega}^2 - \|\phi\|_{-1/2,\Gamma}^2 \lesssim 2a(u, u) + 2(u, u) \\ + (Wu, u) + (V\phi, \phi) \end{aligned}$$

Note that (19) is just the weak formulation of the boundary value problem in Ω_1 if $A(\nabla u_1) \cdot n$ is given on Γ . Now, a coupling can be considered as follows: solve (numerically) one of the equations (14), (15) for $\partial u_2 / \partial n$ in terms of u_2 and insert the resulting expression for $(\partial u_2 / \partial n)|_{\Gamma} = A(\nabla u_1) \cdot n - t_0$ in terms of $u_1|_{\Gamma}$ into (19). Then, (19) has an appropriate form to which FEM can be applied. This works particularly well if a Green's function G for Ω_2 is known. Then, (14) becomes $2(\partial u_2 / \partial n) = -Wu_2$, and inserting this into (19) gives

$$2a(u_1, v) + (Wu_1, v) = 2(t_0, v) + (Wu_0, v) + 2(f, v) \quad (25)$$

Here, the left-hand side satisfies a Gårding inequality in $H^1(\Omega_1)$. Thus, by assuming uniqueness for the solution of the original interface problem, one obtains immediately that every conforming Galerkin method for (25) converges with optimal order. If a Green's function for Ω_2 is not known, then one introduces $\partial u_2 / \partial n$ as an additional unknown. Then, the common direct method (see Johnson and Nedelec, 1980) takes a weak formulation of (15) on the boundary together with (19) and the coupling conditions (3) to form a system that is discretized by approximating u_1 with FE in Ω_1 and $\partial u_2 / \partial n$ with BE on Γ . Another indirect method (see Bielak and MacCamy (1983) and Hsiao (1992)) consists of a single-layer potential ansatz for the exterior problem, that is, the solution of $\Delta u_2 = 0$ in Ω_2 is looked for in the form $u_2 = V\psi$ with an unknown density ψ on Γ . In the last two methods, the resulting matrix is not symmetric and one does not have a Gårding inequality except for linear scalar equations. In the case of

elasticity systems considered in Section 6.1, the operator of the double-layer potential and its adjoint are not compact perturbations in (14) and (15). Thus, standard arguments that guarantee the convergence of the Galerkin method do not apply (cf. Steinbach and Wendland, 2001).

From the error estimate (23), we deduce an $O(h^{1/2})$ convergence of the Galerkin solution for the h-version, in which piecewise linear FE and piecewise constant BE are used, since the solution (u, ϕ) of (21) belongs to $H^{3/2-\epsilon}(\Omega_1) \times H^{-\epsilon}(\Gamma)$ for any $\epsilon > 0$, as follows from the analysis in Costabel (1988a). If $\Gamma = \partial\Omega$ is a smooth manifold, the solution of (21) is also smooth, since the integral operators V, K, K' , and W in (10) to (13) are elliptic pseudodifferential operators. Therefore, for smooth boundaries, the rate of convergence of the Galerkin scheme (23) depends only on the regularity of the FE space X_M and the BE space Y_N . Our symmetric coupling method works for arbitrary meshes (and p-distributions). Especially, we can choose $h_{X_M} = h_{Y_N}$, where h_{X_M} denotes the mesh size of the FE mesh and h_{Y_N} the size of the BE mesh, that is, we can take as grid on the coupling boundary just the mesh points of the FE grid that lie on Γ . Thus, we generalize the results in Wendland (1986, 1988) and Brezzi and Johnson (1979).

Next, we consider the hp-version of the symmetric FE/BE coupling on a geometric mesh. Let Ω be a polygonal domain in \mathbb{R}^2 or a polyhedral domain in \mathbb{R}^3 . Let Ω_h^e be a geometric mesh on Ω , refined appropriately toward vertices of Ω in \mathbb{R}^2 and edges of Ω in \mathbb{R}^3 . Then, Ω_h^e induces a geometric mesh Γ_h^e on the boundary $\Gamma = \partial\Omega$. On the FE mesh Ω_h^e , we approximate u by (affine transformations of) tensor products of antiderivatives of Legendre polynomials together with continuous piecewise linear functions; this gives the space $S^{n,1}(\Omega_h^e)$. On the BE mesh Γ_h^e , we approximate ϕ by elements in $S^{n-1}(\Gamma_h^e)$, which are discontinuous functions and (affine transformations of) tensor products of Legendre polynomials. Then, the above symmetric coupling procedure converges exponentially (see Heuer and Stephan, 1991; Guo and Stephan, 1998).

Theorem 3. Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ be polygonal or polyhedral respectively. Let f, u_0, t_0 be piecewise analytic, then there holds the estimate

$$\|u - u_M\|_{H^1(\Omega)} + \|\phi - \phi_N\|_{H^{-1/2}(\Gamma)} \leq C \begin{cases} (e^{-h_1/2M} + e^{-h_2/2N}), & d = 2 \\ (e^{-h_1/2M} + e^{-h_2/2N}), & d = 3 \end{cases}$$

between the Galerkin solution $u_M \in X_M = S^{n,1}(\Omega_h^e)$, $\phi_N \in Y_N = S^{n-1}(\Gamma_h^e)$, and the exact solution $u, \phi := (\partial u / \partial n)|_\Gamma$ of (21), where the positive constants C, b_1, b_2 are independent of $M = \dim X_M$ and $N = \dim Y_N$.

Next, we comment on the structure of the Galerkin matrix for the symmetric FE/BE coupling procedure, which reflects the saddle-point character of the weak formulation.

To be more specific, let us introduce bases

$$\text{span}\{v_1, \dots, v_M\} = X_M \quad \text{and} \quad \text{span}\{\psi_1, \dots, \psi_N\} = Y_N$$

The basis functions of X_M are supposed to be ordered such that

$$\text{span}\{v_1, \dots, v_{M_0}\} = X_M \cap H_0^{1/2}(\Omega)$$

The basis functions that do not vanish on Γ are then v_{M_0+1}, \dots, v_M . Let us denote the coefficients of u_M and ϕ_N again by u_M and ϕ_N respectively. Further, by u_{M_0} and u_{M-M_0} , we denote the coefficients belonging to the components of u_M that are interior and not interior to Ω , respectively. We obtain a linear system of the form

$$A \begin{pmatrix} u_{M_0} \\ u_{M-M_0} \\ \phi_N \end{pmatrix} = \begin{pmatrix} A & B^T & 0 \\ B & C+W & K^T-I \\ 0 & K-I & -V \end{pmatrix} \begin{pmatrix} u_{M_0} \\ u_{M-M_0} \\ \phi_N \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \quad (26)$$

Here, the block $\begin{pmatrix} A & B^T \\ B & C \end{pmatrix}$ represents a discretization of the FE bilinear form $2a(\cdot, \cdot)$ and corresponds to a Neumann problem, whereas A corresponds to a Dirichlet problem. Note that all the basis functions used to construct A vanish on Γ . The block C deals with the continuous basis functions that are nonzero on Γ . The block W belongs to the same basis functions as C , but is a discretization of the hyper-singular integral operator W . The third diagonal block V only deals with the basis functions of Y_N and belongs to the single-layer potential V . Finally, the blocks I, K , and the transpose of K, K^T , provide the coupling between the two ansatz spaces X_M and Y_N . For I , the bilinear form of the duality between $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$ is used, and K and K^T are discretizations of the double-layer potential K and of its adjoint operator K' . Because of the specific form of A in (26), specific iterative solvers should be chosen for an efficient solution procedure (cf. Section 3) (see Chapter 12, this Volume).

2.1 Convergence analysis

In this section, we prove the existence and uniqueness of the weak (variational) solution of the interface problem (IP) in Section 2 and show the convergence of the Galerkin solution proving Theorem 2 above. The presented analysis follows Carstensen and Stephan (1995a) and uses heavily the strong coerciveness of the Steklov-Poincaré operator (for the exterior problem) and of its discrete analogue.

Firstly, we note that the weak formulation (21) can be rewritten as problem (P).

Problem (P): Find $(u, \phi) \in H^1(\Omega_1) \times H^{-1/2}(\Gamma)$ with

$$B\left(\begin{pmatrix} u \\ \phi \end{pmatrix}, \begin{pmatrix} v \\ \psi \end{pmatrix}\right) = L\left(\begin{pmatrix} v \\ \psi \end{pmatrix}\right) \quad \forall (v, \psi) \in H^1(\Omega_1) \times H^{-1/2}(\Gamma) \quad (27)$$

Here the continuous mapping $B: (H^1(\Omega) \times H^{-1/2}(\Gamma))^2 \rightarrow \mathbb{R}$ and the linear form $L: H^1(\Omega) \times H^{-1/2}(\Gamma) \rightarrow \mathbb{R}$ are defined by

$$B\left(\begin{pmatrix} u \\ \phi \end{pmatrix}, \begin{pmatrix} v \\ \psi \end{pmatrix}\right) := \int_{\Omega_1} A(\nabla u) \cdot \nabla v \, dx + \frac{1}{2} (Wu|_\Gamma - (K' - I)\phi, v|_\Gamma) + \frac{1}{2} (\psi, V\phi + (I - K)u|_\Gamma) \quad (28)$$

$$L\left(\begin{pmatrix} v \\ \psi \end{pmatrix}\right) := \int_{\Omega_1} f \cdot v \, dx + \frac{1}{2} (\psi, (I - K)u_0) + \left(t_0 + \frac{1}{2} Wu_0, v|_\Gamma\right) \quad (29)$$

for any $(u, \phi), (v, \psi) \in H^1(\Omega_1) \times H^{-1/2}(\Gamma)$.

Note that (15) is equivalent to

$$\phi = -V^{-1}(I - K)(u_1 - u_0) \quad (30)$$

which may be used to eliminate $\phi = (\partial u_2 / \partial n)$ in (20). This leads to the following equivalent formulation:

Find $u \in H^1(\Omega_1)$ with

$$A'(u)(\eta) := 2 \int_{\Omega_1} A(\nabla u) \cdot \nabla \eta \, dx + (Su|_\Gamma, \eta|_\Gamma) = L'(\eta) := 2 \int_{\Omega_1} f \cdot \eta \, dx + (2t_0 + Su_0, \eta|_\Gamma) \quad (\eta \in H^1(\Omega)) \quad (31)$$

with the Steklov-Poincaré operator for the exterior problem

$$S := W + (I - K')V^{-1}(I - K) : H^{1/2}(\Gamma) \longrightarrow H^{-1/2}(\Gamma) \quad (32)$$

which is linear, bounded, symmetric, and positive definite.

In the case that A in (IP) is a linear mapping, the following result proves that the bilinear form B satisfies the Babuška-Brezzi condition.

Lemma 2. There exists a constant $\beta > 0$ such that for all $(u, \phi), (v, \psi) \in H^1(\Omega) \times H^{-1/2}(\Gamma)$, we have

$$\beta \cdot \left\| \begin{pmatrix} u-v \\ \phi-\psi \end{pmatrix} \right\|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \cdot \left\| \begin{pmatrix} u-v \\ \phi-\psi \end{pmatrix} \right\|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \leq B\left(\begin{pmatrix} u \\ \phi \end{pmatrix}, \begin{pmatrix} v \\ \psi \end{pmatrix}\right) - B\left(\begin{pmatrix} v \\ \psi \end{pmatrix}, \begin{pmatrix} u \\ \phi \end{pmatrix}\right) \quad (33)$$

with $2\eta := \phi + V^{-1}(I - K)u|_\Gamma$, $2\delta := \psi + V^{-1}(I - K)v|_\Gamma \in H^{-1/2}(\Gamma)$.

Proof. Some calculations show

$$B\left(\begin{pmatrix} u \\ \phi \end{pmatrix}, \begin{pmatrix} v \\ \psi \end{pmatrix}\right) - B\left(\begin{pmatrix} v \\ \psi \end{pmatrix}, \begin{pmatrix} u \\ \phi \end{pmatrix}\right) = \int_{\Omega} ((A \nabla u) - (A \nabla v)) \cdot \nabla (u - v) \, dx + \frac{1}{4} (W(u - v), u - v) + \frac{1}{4} (S(u - v), u - v) + \frac{1}{4} (V(\phi - \psi), \phi - \psi)$$

Since A is uniformly monotone, W is positive semidefinite, S and V are positive definite, the right-hand side is bounded below by $c \left\| \begin{pmatrix} u-v \\ \phi-\psi \end{pmatrix} \right\|_{H^1(\Omega) \times H^{-1/2}(\Gamma)}^2$ with a suitable constant c .

On the other hand, by definition of η, δ , we have with a constant c'

$$\|\eta - \delta\|_{H^{-1/2}(\Gamma)} \leq c' \left\| \begin{pmatrix} u-v \\ \phi-\psi \end{pmatrix} \right\|_{H^1(\Omega) \times H^{-1/2}(\Gamma)}$$

Theorem 4. The interface problem (IP) and the problem (P) have unique solutions.

Proof. The operator A' on the left-hand side in (31) maps $H^1(\Omega_1)$ into its dual; it is continuous, bounded, uniformly monotone, and therefore bijective. This yields the existence of u satisfying (31). Letting ϕ as in (30), we have that (u, ϕ) solves problem (P). Uniqueness of the solution follows from Lemma 2, yielding also the unique solvability of the equivalent interface problem (IP). \square

Next, we treat the discretization of problem (P).

Let $(H_h \times H_h^{-1/2}; h \in I)$ be a family of finite-dimensional subspaces of $H^1(\Omega) \times H^{-1/2}(\Gamma)$. Then, the coupling of FE and BE consists in the following Galerkin procedure.

Definition 1 (Problem (P_h)). For $h \in I$, find $(u_h, \phi_h) \in H_h \times H_h^{-1/2}$ such that

$$B\left(\begin{pmatrix} u_h \\ \phi_h \end{pmatrix}, \begin{pmatrix} v_h \\ \psi_h \end{pmatrix}\right) = L\left(\begin{pmatrix} v_h \\ \psi_h \end{pmatrix}\right) \quad (34)$$

for all $(v_h, \psi_h) \in H_h \times H_h^{-1/2}$.

In order to prove a discrete Babuška-Brezzi condition if A is linear, we need some notations and the positive definiteness of the discrete Steklov-Poincaré operator.

Assumption 1. For any $h \in I$, let $H_h \times H_h^{-1/2} \subseteq H^1(\Omega) \times H^{-1/2}(\Gamma)$, where $I \subseteq (0, 1)$ with $0 \in \bar{I}$, $1 \in H_h^{-1/2}$ for any $h \in I$, where 1 denotes the constant function with value 1.

Let $i_h: H_h \hookrightarrow H^1(\Omega)$ and $j_h: H_h^{-1/2} \hookrightarrow H^{-1/2}(\Gamma)$ denote the canonical injections with their duals $i_h^*: H^1(\Omega)^* \rightarrow H_h^*$ and $j_h^*: H^{-1/2}(\Gamma)^* \rightarrow (H_h^{-1/2})^*$ being projections. Let $\gamma: H^1(\Omega) \rightarrow H^{-1/2}(\Gamma)$ denote the trace operator, $\gamma u = u|_\Gamma$ for all $u \in H^1(\Omega)$, with the dual γ^* . Then, define

$$\begin{aligned} V_h &:= j_h^* V j_h, & K_h &:= j_h^* K j_h \\ W_h &:= i_h^* \gamma^* W \gamma i_h, & K_h' &:= i_h^* \gamma^* K' j_h \end{aligned} \quad (35)$$

and, since V_h is positive definite as well,

$$S_h := W_h - (I_h^* - K_h') V_h^{-1} (I_h - K_h): H_h \rightarrow H_h^* \quad (36)$$

with $I_h := j_h^* \gamma i_h$ and its dual I_h^* .

A key role is played by the following coerciveness of the discrete version of the Steklov–Poincaré operator (see Carstensen and Stephan, 1995a).

Lemma 3. *There exist constants $c_0 > 0$ and $h_0 > 0$ such that for any $h \in I$ with $h < h_0$, we have*

$$(S_h u_h, u_h) \geq c_0 \cdot \|u_h\|_{H^1(\Omega)}^2 \quad \text{for all } u_h \in H_h$$

Lemma 4. *There exist constants $\beta_0 > 0$ and $h_0 > 0$ such that for any $h \in I$ with $h < h_0$, we have that for any $(u_h, \phi_h), (v_h, \psi_h) \in H_h \times H_h^{-1/2}$,*

$$\begin{aligned} \beta_0 \cdot \| \begin{pmatrix} u_h - u_h \\ \phi_h - \phi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \cdot \| \begin{pmatrix} v_h - v_h \\ \psi_h - \psi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \\ \leq B \left(\begin{pmatrix} u_h \\ \phi_h \end{pmatrix}, \begin{pmatrix} v_h \\ \psi_h \end{pmatrix} \right) - B \left(\begin{pmatrix} u_h \\ \phi_h \end{pmatrix}, \begin{pmatrix} v_h - v_h \\ \psi_h - \psi_h \end{pmatrix} \right) \end{aligned} \quad (37)$$

$$\leq B \left(\begin{pmatrix} u_h \\ \phi_h \end{pmatrix}, \begin{pmatrix} v_h \\ \psi_h \end{pmatrix} \right) - B \left(\begin{pmatrix} u_h \\ \phi_h \end{pmatrix}, \begin{pmatrix} v_h - v_h \\ \psi_h - \psi_h \end{pmatrix} \right) \quad (38)$$

with $2\eta_h := \phi_h + V_h^{-1} (I_h - K_h) u_h$, $2\delta_h := \psi_h + V_h^{-1} (I_h - K_h') v_h \in H_h^{-1/2}$. *Proof.* The proof is quite analogous to that of Lemma 2 dealing with the discrete operators (35) and (36). All calculations in the proof of Lemma 2 can be repeated with obvious modifications. Because of Lemma 3, the constants are independent of h as well so that β_0 does not depend on $h < h_0$ (h_0 chosen in Lemma 3). Hence, we may omit the details. \square

Corollary 1. *There exist constants $c_0 > 0$ and $h_0 > 0$ such that for any $h \in I$ with $h < h_0$, the problem (P_h) has a unique solution (u_h, ϕ_h) , and, if (u, ϕ) denotes the solution of (P) , there holds*

$$\begin{aligned} \| \begin{pmatrix} u - u_h \\ \phi - \phi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \\ \leq c_0 \cdot \inf_{\begin{pmatrix} v_h \\ \psi_h \end{pmatrix} \in H_h \times H_h^{-1/2}} \| \begin{pmatrix} u - v_h \\ \phi - \psi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \end{aligned}$$

Proof. The existence and uniqueness of the discrete solutions follows as in the proof of Theorem 4. Let $(U_h, \Phi_h) \in H^h \times H_h^{-1/2}$ be the orthogonal projections onto $H^h \times H_h^{-1/2}$ of the solution (u, ϕ) of problem (P) in $H^1(\Omega) \times H^{-1/2}(\Gamma)$. From Lemma 4, we conclude with appropriate $(\eta_h, \delta_h) \in H^h \times H_h^{-1/2}$ that

$$\begin{aligned} \beta_0 \cdot \| \begin{pmatrix} U_h - u_h \\ \Phi_h - \phi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \cdot \| \begin{pmatrix} U_h - u_h \\ \Phi_h - \phi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \\ \leq B \left(\begin{pmatrix} U_h \\ \Phi_h \end{pmatrix}, \begin{pmatrix} U_h - u_h \\ \Phi_h - \phi_h \end{pmatrix} \right) - B \left(\begin{pmatrix} U_h \\ \Phi_h \end{pmatrix}, \begin{pmatrix} U_h - u_h \\ \Phi_h - \phi_h \end{pmatrix} \right) \end{aligned}$$

Using the Galerkin conditions and the Lipschitz continuity of B , with related constant L (which follows since A is Lipschitz continuous), we get that the right-hand side is bounded by

$$L \cdot \| \begin{pmatrix} U_h - u_h \\ \Phi_h - \phi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \cdot \| \begin{pmatrix} U_h - u_h \\ \Phi_h - \phi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)}$$

Dividing the whole estimate by $\| \begin{pmatrix} U_h - u_h \\ \Phi_h - \phi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)}$ proves

$$\| \begin{pmatrix} U_h - u_h \\ \Phi_h - \phi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \leq \frac{L}{\beta_0} \cdot \| \begin{pmatrix} U_h - u_h \\ \Phi_h - \phi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)}$$

From this, the triangle inequality yields the assertion. \square

2.2 Adaptive FE/BE coupling

In this section, we present a posteriori error estimates for the h-version of the symmetric coupling method.

For the efficiency of FE and BE computations, it is of high importance to use *local* rather than global mesh refinement in order to keep the number of unknowns reasonably small. Often, in particular for nonlinear problems, any a priori information about the solution that would allow the construction of a suitably refined mesh is not available. Thus, one has to estimate the error a posteriori.

For the FEM, there are some well-known error estimators:

1. Inserting the computed solution into the partial differential equation, one obtains a *residual*. For an *explicit residual error indicator*, a suitable norm of the residual is computed; this also involves jump terms at the element boundaries.
2. Error estimators based on *hierarchical bases* are often used in the context of multilevel methods that allow fast iterative solution procedures for large systems of equations. For an overview on the estimators, see Bank and Smith (1993).
3. To obtain an approximation of the error, one may solve *local problems* on each FE or on patches of FE (see Bank and Weiser, 1985). The right-hand side of the

local problems involve the residual. So these estimators are called *implicit residual estimators*.

If δ is the estimated error and e is the actual error, the ratio δ/e is called the *effectivity index*. The advantage of the third approach is that one can expect a good effectivity index. For linear elliptic problems with a positive definite variational formulation, the solution of infinite-dimensional local Neumann problems yields an upper bound on the energy norm of the error in the entire domain; there are no multiplicative constants in this estimate, and the effectivity often is not much larger than 1. In practice, the local problems are solved approximately by using a higher polynomial degree or a finer mesh. In contrast, the first approach merely yields a bound up to a multiplicative constant that is difficult to determine. Most often, this method is just used as a mesh refinement criterion. Of course, the evaluation of this error indicator is much cheaper than the solution of a local problem. Explicit residual error indicators can be carried over to boundary element methods; inserting the approximate solution into the integral equations, one obtains a residual, which has to be evaluated in a suitable norm (see Subsection 2.2.1). It is also possible to establish a hierarchical basis estimator for BE (see Subsection 2.2.2). The use of error estimators based on local problems for the FE/BE coupling is given in Subsection 2.2.3 (see Chapter 4, this Volume, Chapter 2, Volume 2).

2.2.1 Residual based error indicators

In this section, we present an a posteriori error estimate from Carstensen and Stephan (1995a). For simplicity, we restrict ourselves to linear functions on triangles as FE in H_h and to piecewise constant functions in $H_h^{-1/2}$.

Assumption 2. Let Ω be a two-dimensional domain with polygonal boundary Γ on which we consider a family $\mathcal{T} := \{T_i; i \in I\}$ of decompositions $T_i = \{\Delta_1, \dots, \Delta_N\}$ of Ω in closed triangles $\Delta_1, \dots, \Delta_N$ such that $\Omega = \bigcup_{i \in I} \Delta_i$, and two different triangles are disjoint or have a side or a vertex in common. Let S_h denote the sides, that is,

$$S_h = \{ \partial T_i \cap \partial T_j : i \neq j \text{ with } \partial T_i \cap \partial T_j \text{ as the common side} \}$$

∂T_i being the boundary of T_i . Let

$$\mathcal{G}_h = \{E : E \in S_h \text{ with } E \subseteq \Gamma\}$$

be the set of ‘boundary sides’ and let

$$\mathcal{S}_h^0 = S_h \setminus \mathcal{G}_h$$

be the set of ‘interior sides’.

We assume that all the angles of some $\Delta \in \mathcal{T}_h \in \mathcal{T}$ are $\geq \Theta$ for some fixed $\Theta > 0$, which does not depend on Δ or \mathcal{T}_h .

Then define

$$\begin{aligned} H_h &:= \{ \eta_h \in C(\Omega) : \eta_h|_\Delta \in P_1 \text{ for any } \Delta \in \mathcal{T}_h \} \quad (39) \\ H_h^{-1/2} &:= \{ \eta_h \in L^\infty(\Gamma) : \eta_h|_E \in P_0 \text{ for any } E \in \mathcal{G}_h \} \end{aligned} \quad (40)$$

where P_j denotes the polynomials with degree $\leq j$.

For fixed \mathcal{T}_h , let h be the piecewise constant function defined such that the constants $h|_\Delta$ and $h|_E$ equal the element sizes $\text{diam}(\Delta)$ of $\Delta \in \mathcal{T}_h$ and $\text{diam}(E)$ of $E \in \mathcal{G}_h$.

We assume that $A(\nabla v_h) \in C^1(\Delta)$ for any $\Delta \in \mathcal{T}_h \in \mathcal{T}$ and any trial function $v_h \in H_h$. Finally, let $f \in L^2(\Omega)$, $u_0 \in H^1(\Gamma)$, and $t_0 \in L^2(\Gamma)$.

Let n be the exterior normal on Γ , and on any element boundary $\partial\Delta$, let n have a fixed orientation so that $[A(\nabla u_h) \cdot n]_E \in L^2(E)$ denotes the jump of the discrete tractions $A(\nabla u_h) \cdot n$ over the side $E \in \mathcal{G}_h$. Define

$$R_1^2 := \sum_{\Delta \in \mathcal{T}_h} \text{diam}(\Delta)^2 \cdot \int_\Delta |f + \text{div } A(\nabla u_h)|^2 dx \quad (41)$$

$$R_2^2 := \sum_{E \in \mathcal{G}_h^0} \text{diam}(E) \cdot \int_E |[A(\nabla u_h) \cdot n]|^2 ds \quad (42)$$

$$\begin{aligned} R_3 &:= \left\| \sqrt{h} \cdot \left(t_0 - A(\nabla u_h) \cdot n + \frac{1}{2} W(u_0 - u_h)|_\Gamma \right) \right. \\ &\quad \left. - \frac{1}{2} (K' - I) \phi_h \right\|_{L^2(\Gamma)} \end{aligned} \quad (43)$$

$$\begin{aligned} R_4 &:= \sum_{E \in \mathcal{G}_h} \text{diam}(E)^{1/2} \cdot \left\| \frac{\partial}{\partial s} ((I - K)(u_0 - u_h)|_\Gamma) \right. \\ &\quad \left. - V \phi_h \right\|_{L^2(E)} \end{aligned} \quad (44)$$

Under the above assumptions and notations, there holds the following a posteriori estimate, in which (u, ϕ) and (u_h, ϕ_h) solve problem (P) and (P_h) respectively (see Carstensen and Stephan, 1995a).

Theorem 5. *There exists some constant $c > 0$ such that for any $h \in I$ with $h < h_0$ (h_0 from Lemma 3), we have*

$$\| \begin{pmatrix} u - u_h \\ \phi - \phi_h \end{pmatrix} \|_{H^1(\Omega) \times H^{-1/2}(\Gamma)} \leq c \cdot (R_1 + R_2 + R_3 + R_4)$$

Note that R_1, \dots, R_4 can be computed (at least numerically) as far as the solution (u_h, ϕ_h) of problem (P_h) is known. The proof of Theorem 5 is based on Lemma 4. A corresponding adaptive feedback procedure is described in

Carstensen and Stephan (1995a) and is extended to elasticity problems in Carstensen, Funken and Stephan (1997) and to interface problems with viscoplastic and plastic material in Carstensen, Zarrabi and Stephan (1996). A posteriori error estimates with lower bounds yielding efficiency are given in Carstensen (1996b) for uniform meshes.

2.2.2 Adaptive FE/BE coupling with a Schur complement error indicator

Recently, the use of adaptive hierarchical methods has been becoming increasingly popular. Using the discretization of the Steklov–Poincaré operator, we present for the symmetric FE/BE coupling method an a posteriori error estimate with 'local' error indicators; for an alternative method that uses the full coupling formulation, see Mund and Stephan (1999). By using stable hierarchical basis decompositions for FE (see Yserentant, 1986), we derive two-level subspace decompositions for locally refined meshes. Assuming a saturation condition to hold, as mentioned in Krebs, Maischak and Stephan (2001), an adaptive algorithm is formulated to compute the FE solution on a sequence of refined meshes in the interior domain and on the interface boundary. At the end of this subsection, we present numerical experiments that show the efficiency and reliability of the error indicators.

Let $\rho \in C^1(\mathbb{R}_+)$ satisfy the conditions

$$\rho_0 \leq \rho(t) \leq \rho_1 \quad \text{and} \quad \rho_2 \leq \rho(t) + t\rho'(t) \leq \rho_3 \quad (45)$$

for some global constants $\rho_0, \rho_1, \rho_2, \rho_3 > 0$. We consider the following nonlinear interface problem (NP) (cf. Gatica and Hsiao, 1989) in \mathbb{R}^2 :

Problem (NP): Given the functions $f: \Omega_1 \rightarrow \mathbb{R}$ and $u_0, i_0: \Gamma \rightarrow \mathbb{R}$, find $u_i: \Omega_i \rightarrow \mathbb{R}, i = 1, 2$, and $b \in \mathbb{R}$ such that

$$-\operatorname{div}(\rho(|\nabla u_1|) \nabla u_1) = f \quad \text{in } \Omega_1 \quad (46a)$$

$$-\Delta u_2 = 0 \quad \text{in } \Omega_2 \quad (46b)$$

$$u_1 - u_2 = u_0 \quad \text{on } \Gamma \quad (46c)$$

$$\rho(|\nabla u_1|) \frac{\partial u_1}{\partial n} - \frac{\partial u_2}{\partial n} = i_0 \quad \text{on } \Gamma \quad (46d)$$

$$u_2(x) = b \log |x| + o(1) \quad \text{for } |x| \rightarrow \infty \quad (46e)$$

where $\partial v / \partial n$ is the normal derivative of v pointing from Ω_1 into Ω_2 .

By a symmetric coupling method (Costabel, 1988b), the problem (46) is transformed into the following variational problem (cf. Carstensen and Stephan, 1995a; Stephan, 1992).

Given $f \in (H^1(\Omega_1))'$, $u_0 \in H^{1/2}(\Gamma)$, and $i_0 \in H^{-1/2}(\Gamma)$, find $u \in H^1(\Omega_1)$ and $\phi \in H^{-1/2}(\Gamma)$ such that

$$a(u, v) + B(u, \phi; v, \psi) = \mathcal{L}(v, \psi) \quad (47)$$

for all $v \in H^1(\Omega_1)$ and $\psi \in H^{-1/2}(\Gamma)$, where the form $a(\cdot, \cdot)$ is defined as

$$a(u, v) := 2 \int_{\Omega_1} \rho(|\nabla u|) \nabla u \nabla v \, dx$$

the bilinear form $B(\cdot, \cdot)$ is defined as

$$B(u, \phi; v, \psi) := (Wu|_{\Gamma} + (K' - I)\phi, v|_{\Gamma}) - (\psi, (K - I)u|_{\Gamma} - V\phi)$$

and the linear form $\mathcal{L}(\cdot)$ is defined as

$$\mathcal{L}(v, \psi) := 2(f, v) + (2i_0 + Wu_0, v|_{\Gamma}) - (\psi, (K - I)u_0)$$

Here, (\cdot, \cdot) and $\langle \cdot, \cdot \rangle$ denote the duality pairings between $(H^1(\Omega_1))'$ and $H^1(\Omega_1)$ and between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$ respectively. The unknowns in (46) satisfy $u_1 = u$ and $\partial u_2 / \partial n = \phi$ and u_2 can be obtained via a representation formula (Costabel, 1988b).

Lemma 5. The following problem is equivalent to (47).

Find $u \in H^1(\Omega_1)$ such that

$$a(u, v) + (Su|_{\Gamma}, v) = F(v) \quad \forall v \in H^1(\Omega_1) \quad (48)$$

where

$$F(v) := 2 \int_{\Omega_1} f v \, dx + (2i_0 + Su_0, v|_{\Gamma})$$

$a(\cdot, \cdot)$ as in (47), and the Steklov–Poincaré operator for the exterior domain $S := W + (K' - I)V^{-1}(K - I)$ is a continuous map from $H^{1/2}(\Gamma)$ into $H^{-1/2}(\Gamma)$.

Firstly, we describe the coupling of the FEM and the BEM to compute approximations to the solution (u, ϕ) of (47). For this purpose, we consider regular triangulations ω_H of Ω_1 and partitions γ_H of Γ . Our test and trial spaces are defined as

$$T_H := \{v_H: \Omega_1 \rightarrow \mathbb{R}; v_H \text{ p.w. linear on } \omega_H, \quad (49)$$

$$v_H \in C^0(\Omega_1)\},$$

$$\tau_H := \{\psi_H: \Gamma \rightarrow \mathbb{R}; \psi_H \text{ p.w. constant on } \gamma_H\} \quad (50)$$

We assume that the mesh for the discretization of the BE part γ_H is induced by that of the finite element part. This yields the following discretization of problem (47).

Find $(u_H, \phi_H) \in T_H \times \tau_H$ such that

$$2 \int_{\Omega_1} \rho(|\nabla u_H|) \nabla u_H \nabla v \, dx + B(u_H, \phi_H; v, \psi) = \mathcal{L}(v, \psi) \quad (51)$$

for all $(v, \psi) \in T_H \times \tau_H$.

The application of Newton's method to (51) yields a sequence of linear systems to be solved. Given an initial guess $(u_H^{(0)}, \phi_H^{(0)})$, we seek to find

$$(u_H^{(j)}, \phi_H^{(j)}) = (u_H^{(j-1)}, \phi_H^{(j-1)}) + (d_H^{(j)}, \delta_H^{(j)}) \quad (j = 1, 2, \dots)$$

such that

$$a_{u_H^{(j-1)}}(d_H^{(j)}, v) + B(d_H^{(j)}, \delta_H^{(j)}; v, \psi) = \mathcal{L}(v, \psi) - a(u_H^{(j-1)}, v) - B(u_H^{(j-1)}, \phi_H^{(j-1)}; v, \psi) \quad (52)$$

for all $(v, \psi) \in T_H \times \tau_H$ with $a(\cdot, \cdot)$, $B(\cdot, \cdot; \cdot, \cdot)$, and $\mathcal{L}(\cdot, \cdot)$ as in (47). The bilinear form $a_{u_H^{(j-1)}}(\cdot, \cdot)$ is defined by

$$a_{u_H^{(j-1)}}(u, v) := 2 \int_{\Omega_1} (\bar{\rho}(|\nabla u|) \nabla u) \nabla v \, dx \quad (53)$$

and $\bar{\rho} \in \mathbb{R}^2$ is the Jacobian of $x \rightarrow \rho(|x|)x$, that is,

$$\bar{\rho} = \rho(|x|)I_{2 \times 2} + \rho'(|x|) \frac{x \cdot x^T}{|x|} \quad (x \in \mathbb{R}^2)$$

From the assumptions on ρ in (45), it follows that there exist constants $\nu, \mu > 0$ such that

$$a_{u_H^{(j-1)}}(u, v) \leq \nu \|u\|_{H^1(\Omega_1)} \|v\|_{H^1(\Omega_1)}$$

and

$$\mu \|u\|_{H^1(\Omega_1)}^2 \leq a_{u_H^{(j-1)}}(u, u) \quad (54)$$

for all $u, v \in H^1(\Omega_1)$.

Since ρ is sufficiently smooth, the energy functional of (47) is strictly convex, and hence, Newton's method converges locally.

For the implementation of (52), we define the piecewise linear basis functions

$$b_i(v_j) := \delta_{i,j} \quad (1 \leq j \leq n_{\text{in}}, \quad n_{\text{in}} + 1 \leq j \leq n_T)$$

where $v_i \in \Omega_1 \setminus \Gamma$ ($1 \leq i \leq n_{\text{in}}$) are the inner nodes of ω_H and $v_i \in \Gamma$ ($n_{\text{in}} + 1 \leq i \leq n_T := \dim T_H$) are the boundary nodes of ω_H counted along the closed curve Γ .

On the boundary, the following basis of τ_H is introduced. Let $\mu_i \in \gamma_H$ be the BE induced by the nodes $v_{n_{\text{in}}+1}, v_{n_{\text{in}}+2}, \dots, v_{n_T}$ ($1 \leq i \leq n_T - 1, n_T := \dim \tau_H$) and μ_{n_T} by the nodes $v_{n_T}, v_{n_{\text{in}}+1}$. With each μ_i , we associate the basis

function

$$\beta_i(x) := \begin{cases} 1 & \text{if } x \in \mu_i \\ 0 & \text{if } x \in \Gamma \setminus \mu_i \end{cases}$$

With the basis functions b_i and β_i , (52) yields a linear system, which may be solved with the hybrid modified conjugate residual (HMCRCR) scheme together with efficient preconditioners (Hahne *et al.*; Heuer, Maischak and Stephan, 1999; Mund and Stephan, 1997).

In Mund and Stephan (1999), an adaptive algorithm is given on the basis of a posteriori error estimates of the solution (u_H, ϕ_H) of (51). Here, we apply a Schur complement method based on a Galerkin discretization of the variational formulation (48), eliminating the unknown vector ϕ . In this way, we also obtain a discretization of the Steklov–Poincaré operator, which will be used to develop an a posteriori error indicator that needs only a refinement of the mesh defining T_H and does not need a finer discretization as τ_H .

Next, we introduce hierarchical two-level decompositions for the finite element space T_h on ω_h (cf. (49)), where we get ω_h by the refinement shown in Figure 1.

These decompositions will be used to derive an a posteriori error estimate for the Galerkin solution to (47), which is obtained by applying a Schur complement to (51).

We take the hierarchical two-level subspace decomposition

$$T_h := T_H \oplus L_h, \quad L_h := T_1 \oplus T_2 \oplus \dots \oplus T_n$$

with $T_i := \operatorname{span}\{\hat{b}_i\}$, where \hat{b}_i denotes the piecewise linear basis functions in the new n node points v_i of the fine grid (Yserentant, 1986; Mund and Stephan, 1999). Let $P_H: T_h \rightarrow T_H, P_L: T_h \rightarrow T_L$ be the Galerkin projections with respect to the bilinear form $B(\cdot, \cdot)$, which is defined as

$$b(u, v) := \int_{\Omega_1} (\nabla u \nabla v + uv) \, dx \quad (55)$$

For all $u \in T_h$, we define P_H and P_L by

$$b(P_H u, v) = b(u, v) \quad \forall v \in T_H$$

$$b(P_L u, v) = b(u, v) \quad \forall v \in T_L$$

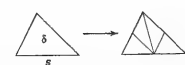


Figure 1. Refinement of $\delta \in \omega_H$. The longest edge of δ is denoted by s . The new nodes are the midpoints of the edges of δ .

Now, we introduce the approximate Steklov–Poincaré operator on fine-mesh functions

$$\tilde{S}_h := W_h + (K_{H,h}^* - I_{H,h}^*)V_H^{-1}(K_{H,h} - I_{H,h}) \quad (56)$$

where, for $u, v \in T_h$ and $\phi, \psi \in \mathcal{V}_H$,

$$\begin{aligned} \langle W_h u|_{\Gamma}, v|_{\Gamma} \rangle &= \langle W u|_{\Gamma}, v|_{\Gamma} \rangle \\ \langle (K_{H,h} - I_{H,h})u|_{\Gamma}, \psi|_{\Gamma} \rangle &= \langle (K - I)u|_{\Gamma}, \psi|_{\Gamma} \rangle \\ \langle V_H \phi|_{\Gamma}, \psi|_{\Gamma} \rangle &= \langle V \phi|_{\Gamma}, \psi|_{\Gamma} \rangle \\ \langle (K_{H,h}^* - I_{H,h}^*)\phi|_{\Gamma}, v|_{\Gamma} \rangle &= \langle (K^* - I)\phi|_{\Gamma}, v|_{\Gamma} \rangle \end{aligned}$$

Furthermore, we consider the discrete Steklov–Poincaré operator

$$S_H := W_H + (K_{H,H}^* - I_{H,H}^*)V_H^{-1}(K_{H,H} - I_{H,H}) \quad (57)$$

on coarse mesh functions, in which the operators are defined as above by substituting T_H for T_h .

With the discrete Steklov–Poincaré operators \tilde{S}_h and S_H , we formulate discrete problems to (48):

Find $u_H \in T_H$ such that

$$a(u_H, v) + \langle S_H u_H|_{\Gamma}, v|_{\Gamma} \rangle = F_H(v) \quad \forall v \in T_H \quad (58)$$

and

find $\tilde{u}_h \in T_h$ such that

$$a(\tilde{u}_h, v) + \langle \tilde{S}_h \tilde{u}_h|_{\Gamma}, v|_{\Gamma} \rangle = \tilde{F}_h(v) \quad \forall v \in T_h \quad (59)$$

where $F_H(\cdot)$ and $\tilde{F}_h(\cdot)$ are obtained by substituting S_H for S in F of (48) and \tilde{S}_h respectively.

For our analysis, to derive an a posteriori error estimate (Theorem 6), we have to make the following saturation assumption.

Assumption 3. Let u, u_H, \tilde{u}_h be defined as in (48), (58), and (59). There exists a constant $\kappa \in (0, 1)$ independent of H, h such that

$$\|u - \tilde{u}_h\|_{H^1(\Omega)} \leq \kappa \|u - u_H\|_{H^1(\Omega)}$$

The following a posteriori error estimate is proved in Krebs, Maischak and Stephan (2001).

Theorem 6. Assume that the above saturation assumption holds. Let $T_0 \subset T_1 \subset T_2 \subset \dots$ be a sequence of hierarchical subspaces, where T_0 is an initial FEM space (cf. (49)). The refinement of all triangles defining T_k according to Figure 1 gives us T_{k+1} . Let k denote the number of the refinement level

and u_k the corresponding Galerkin solution of (58) and u the exact solution of (48), then there are constants $\zeta_1, \zeta_2 > 0$, $k_0 \in \mathbb{N}_0$, such that for all $k \geq k_0$,

$$\zeta_1 \left(\sum_{i=1}^n \tilde{\theta}_{i,k}^2 \right)^{1/2} \leq \|u - u_k\|_{H^1(\Omega)} \leq \zeta_2 \left(\sum_{i=1}^n \tilde{\theta}_{i,k}^2 \right)^{1/2} \quad (60)$$

where the local error indicators

$$\tilde{\theta}_{i,k} := \frac{|2\tilde{\theta}_G(b_{i,k}) + \partial_F(b_{i,k})|}{\|b_{i,k}\|_{H^1(\Omega_i)}} \quad (61)$$

are obtained via the basis functions $b_{i,k} \in T_{h,k} \setminus T_k$ by a domain part

$$\tilde{\theta}_G(b_{i,k}) := \int_{\Omega_i} f b_{i,k} dx - \int_{\Omega_i} \rho(|\nabla u_k|) \nabla u_k \nabla b_{i,k} dx \quad (62)$$

and a boundary part

$$\partial_F(b_{i,k}) := (2\tilde{\theta}_0 + \tilde{S}_{h,k} u_0, b_{i,k}|_{\Gamma}) - (\tilde{S}_{h,k} u_{k,\Gamma}, b_{i,k}|_{\Gamma}) \quad (63)$$

with $\tilde{S}_{h,k}$ defined as in (56) with respect to $T_{h,k}$. T_k instead of T_h .

Next, we list the numerical experiment from Krebs, Maischak and Stephan (2001) for (NP) with $Q = 1$ and choose Ω_1 to be the L-shaped domain with corners at $(0, 0)$, $(0, 1/4)$, $(-1/4, 1/4)$, $(-1/4, -1/4)$, $(1/4, -1/4)$, $(1/4, 0)$. The exact solution of the model problem (NP) is given by

$$\begin{aligned} u_1(r, \alpha) &= r^{2/3} \sin \frac{2}{3} \left(\alpha - \frac{\pi}{2} \right) \\ u_2(x_1, x_2) &= \log \sqrt{\left(x_1 + \frac{1}{8} \right)^2 + \left(x_2 + \frac{1}{8} \right)^2} \quad (64) \end{aligned}$$

The functions u_0, t_0, f are chosen to yield the exact solution. The quantities in Table 1 are given as follows: With k we denote the refinement level, with n_k the total number of unknowns, and with N_k the total number of triangles defining T_k . The error E_k is defined as

$$E_k := \|u - u_k\|_{1,\Omega}$$

The global error indicator η_k is defined by

$$\begin{aligned} \eta_k &= \left(\sum_{i=1}^{N_k} \tilde{\eta}_{i,k}^2 \right)^{1/2} \\ \tilde{\eta}_{i,k} &:= (\tilde{\theta}_{i,k}^2 + \tilde{\theta}_{\alpha,k}^2 + \tilde{\theta}_{\beta,k}^2)^{1/2} \quad (i = 1, \dots, N_k) \end{aligned}$$

Table 1. Results for adaptive algorithm based on Theorem 6 for (NP) with u_1, u_2 from (64), $\zeta = 0.15$.

| L | n_k | $\dim T_k$ | $\dim \mathcal{V}_k$ | E_k | η_k | η_k/E_k | κ_k | α_k |
|-----|-------|------------|----------------------|---------|----------|--------------|------------|------------|
| 0 | 37 | 21 | 16 | 0.10608 | 0.13067 | 1.232 | — | — |
| 1 | 55 | 37 | 18 | 0.07596 | 0.08283 | 1.090 | 0.716 | 0.842 |
| 2 | 78 | 58 | 20 | 0.05511 | 0.06495 | 1.179 | 0.725 | 0.919 |
| 3 | 109 | 85 | 24 | 0.04510 | 0.05596 | 1.241 | 0.818 | 0.599 |
| 4 | 163 | 129 | 34 | 0.03626 | 0.04373 | 1.206 | 0.804 | 0.542 |
| 5 | 454 | 396 | 58 | 0.02063 | 0.02419 | 1.172 | 0.569 | 0.550 |
| 6 | 677 | 595 | 82 | 0.01654 | 0.01936 | 1.171 | 0.802 | 0.554 |
| 7 | 1972 | 1840 | 132 | 0.01008 | 0.01108 | 1.100 | 0.609 | 0.464 |

Here i_1, i_2, i_3 denote the three edges and the corresponding new base functions for every element of the old mesh. The values of the quotient η_k/E_k , the efficiency index, indicate the efficiency of the error indicator η_k and confirm Theorem 6. The quantity

$$\kappa_k := \frac{\|u - u_k\|_{1,\Omega_1}}{\|u - u_{k+1}\|_{1,\Omega_1}}$$

estimates the saturation constant κ . Since κ_k is bounded by a constant less than 1, the saturation condition is satisfied for the sequence of meshes that is generated by our adaptive algorithm. With the above value for the steering parameter ζ , we obtain

$$\kappa_k \approx 1.45^{-0.5} \approx 0.83$$

which, in view of Table 1, turns out to be an upper bound for the κ_k . The experimental convergences rates α_k are given by

$$\alpha_k = \frac{\log(E_k/E_{k-1})}{\log(n_{k-1}/n_k)}$$

From Table 1, we see that α_k approaches $1/2$, which is the convergence rate in case of a smooth solution, thus showing the quality of the adaptive algorithm. The above hierarchical method is easily implemented, since for the computation of the error indicators one can use the same routine as for the computation of the entries of the Galerkin matrix.

2.2.3 The implicit residual error estimator

Error estimators based on local problems cannot directly be carried over to BE because of the nonlocal nature of the integral operators. In Brink and Stephan (1999), we combine the technique of solving local problems in the FEM domain with an explicit residual error estimator for one equation of the BE formulation. For each FE, a

Neumann problem is solved. As in the pure FEM case, the boundary data of each local problem has to satisfy an equilibration condition that ensures solvability.

For simplicity, let us consider the interface problem (IP) with $d = 3$ and $t_0 = u_0 = 0$ and assume that the domain Ω is partitioned as $\Omega = \cup \{T; T \in T_h\}$. The elements T typically are open tetrahedra. For $T \neq T'$, $\bar{T} \cap \bar{T}'$ is either empty or a common vertex or edge or face.

Let $(u_h, \phi_h) \in H^1(\Omega) \times H^{-1/2}(\Gamma)$ denote a computed approximation of the solution of (21). (u_h, ϕ_h) may contain errors because of an approximate solution of an appropriate discrete system. In view of (14),

$$2\tilde{\phi}_h := (I - K^*)\phi_h - Wu_h$$

is an approximation of the normal derivative ϕ on Γ .

We need a continuous symmetric bilinear form $\tilde{a}(\cdot, \cdot)$ on $H^1(\Omega)$. For the time being, \tilde{a} need not be specified further. The restriction to an element T is denoted by \hat{a}_T and is such that

$$\tilde{a}(w, v) = \sum_{T \in \mathcal{T}_h} \hat{a}_T(w|_T, v|_T) \quad \forall w, v \in H^1(\Omega) \quad (65)$$

For every element, we set

$$\mathcal{W}_T := \{v|_T; v \in H^1(\Omega)\} \quad (66)$$

The kernel of \hat{a}_T is

$$\mathcal{Z}_T := \{v \in \mathcal{W}_T; \hat{a}_T(w, v) = 0 \quad \forall w \in \mathcal{W}_T\} \quad (67)$$

The bilinear form \tilde{a} is required to be $H^1(\Omega)$ -elliptic, that is, there exists a positive constant α such that

$$\tilde{a}(v, v) \geq \alpha \|v\|_{1,\Omega}^2 \quad \forall v \in H^1(\Omega) \quad (68)$$

Similarly, for all elements, \hat{a}_T is required to be $\mathcal{W}_T/\mathcal{Z}_T$ -elliptic.

For the error estimator, the following local problem with Neumann boundary conditions is solved for each element. Find $w_T \in \mathcal{W}_T$ such that

$$\hat{a}_T(w_T, v) = \ell_T(v) \quad \forall v \in \mathcal{W}_T \quad (69)$$

with the linear functional

$$\ell_T(v) := 2 \int_T f v dx - 2 \int_T A(\nabla u_h) \cdot \nabla v dx + 2 \int_{\partial T} q_T v ds \quad (70)$$

The functions $q_T \in L^2(\partial T)$ are prescribed normal derivatives. In Brink and Stephan (1999), we comment on how to obtain suitable q_T . Hence, we only assume that the following conditions are satisfied:

1. The boundary conditions on Γ_N are

$$q_T = g \quad \text{on } \partial T \cap \Gamma_N \quad (71)$$
2. The boundary conditions on the coupling interface are

$$q_T = \phi_h \quad \text{on } \partial T \cap \Gamma \quad (72)$$
3. On all element boundaries in the interior of Ω , there holds

$$q_T = -q_{T'} \quad \text{on } \bar{T} \cap \bar{T}' \quad (T \neq T') \quad (73)$$
4. For all elements, there holds the equilibration condition

$$\ell_T(v) = 0 \quad \forall v \in Z_T \quad (74)$$

This condition is obviously necessary for the solvability of (69). For completeness, we prove the following a posteriori error estimate from Brink and Stephan (1999).

Theorem 7. Let bilinear forms \hat{a} and \hat{a}_T be $H^1(\Omega)$ -elliptic and \mathcal{W}_T/Z_T -elliptic respectively. Assume that the conditions (71) to (74) hold. Then

$$\|u_1 - u_h\|_{1,\Omega} + \|\phi - \phi_h\|_{-1/2,\Gamma} \leq C \|V\phi_h + (I - K)u_h\|_{1/2,\Gamma} + C \left\{ \sum_{T \in \mathcal{T}_h} \hat{a}_T(w_T, w_T) \right\}^{1/2}$$

where w_T are the solutions of the local problems (69).

Proof. Owing to the ellipticity of \hat{a}_T , the local problems (69) are solvable. We define $z \in H^1(\Omega)$ to be the unique solution of

$$\hat{a}(z, v) = 2 \int_{\Omega} \{A(\nabla u_1) - A(\nabla u_h)\} \cdot \nabla v \, dx - 2(v, \phi - \phi_h) \quad \forall v \in H^1(\Omega) \quad (75)$$

Note that

$$\begin{aligned} \hat{a}(z, v) &= 2 \int_{\Omega} \{A(\nabla u_1) - A(\nabla u_h)\} \cdot \nabla v \, dx \\ &\quad + (v, W(u_1 - u_h) - (I - K')(\phi - \phi_h)) \\ &= -2 \int_{\Omega} A(\nabla u_h) \cdot \nabla v \, dx \\ &\quad + 2(v, \tilde{\phi}_h) + 2 \int_{\Omega} f v \, dx \end{aligned} \quad (76)$$

by (21). From the uniform monotonicity of A , see (6), and the positivity of the integral operators V and W , we

conclude,

$$\begin{aligned} &\alpha(\|u_1 - u_h\|_{1,\Omega}^2 + \|\phi - \phi_h\|_{-1/2,\Gamma}^2) \\ &\leq 2 \int_{\Omega} \{A(\nabla u_1) - A(\nabla u_h)\} \cdot \nabla(u_1 - u_h) \, dx \\ &\quad + (u_1 - u_h, W(u_1 - u_h)) + (\phi - \phi_h, V(\phi - \phi_h)) \\ &= 2 \int_{\Omega} \{A(\nabla u_1) - A(\nabla u_h)\} \cdot \nabla(u_1 - u_h) \, dx \\ &\quad + (u_1 - u_h, W(u_1 - u_h)) - (u_1 - u_h, (I - K')(\phi - \phi_h)) \\ &\quad + (\phi - \phi_h, V(\phi - \phi_h)) + (\phi - \phi_h, (I - K)(u_1 - u_h)) \\ &= \hat{a}(z, u_1 - u_h) + \left(\phi - \phi_h, -V\phi_h - \left(\frac{1}{2}I - K \right) u_h \right) \end{aligned}$$

where we used integral equation (15) for the exact solution. Exploiting the continuity of $\hat{a}(\cdot, \cdot)$ and (\cdot, \cdot) , we obtain

$$\begin{aligned} &\|u_1 - u_h\|_{1,\Omega} + \|\phi - \phi_h\|_{-1/2,\Gamma} \\ &\leq C(\|z\|_{1,\Omega} + \|V\phi_h + (I - K)u_h\|_{1/2,\Gamma}) \end{aligned} \quad (77)$$

Because of (76), solving (75) corresponds to minimizing

$$\begin{aligned} \Phi(v) &:= \frac{1}{2} \hat{a}(v, v) + 2 \int_{\Omega} A(\nabla u_h) \\ &\quad \cdot \nabla v \, dx - 2(v, \tilde{\phi}_h) - 2 \int_{\Omega} f v \, dx \end{aligned} \quad (78)$$

over $H^1(\Omega)$, and

$$\inf_{v \in H^1(\Omega)} \Phi(v) = \Phi(z) = \frac{1}{2} \hat{a}(z, z) \quad (79)$$

Let us define

$$\Phi_T(v) := \frac{1}{2} \hat{a}_T(v, v) - \ell_T(v) \quad (80)$$

on \mathcal{W}_T . Then, by the trace theorem and (73),

$$\Phi(v) = \sum_{T \in \mathcal{T}_h} \Phi_T(v) \quad \forall v \in H^1(\Omega)$$

and

$$\begin{aligned} \inf_{v \in H^1(\Omega)} \Phi(v) &\geq \inf_{v \in L^2(\Omega), \forall T \in \mathcal{T}_h, \forall T} \inf_{v \in \mathcal{W}_T} \sum_{T \in \mathcal{T}_h} \Phi_T(v) \\ &= \sum_{T \in \mathcal{T}_h} \inf_{v \in \mathcal{W}_T} \Phi_T(v) \end{aligned}$$

The elementwise minimizers on the right-hand side are the solutions of local problems (69) and thus (79),

$$-\frac{1}{2} \hat{a}(z, z) \geq \sum_{T \in \mathcal{T}_h} -\frac{1}{2} \hat{a}_T(w_T, w_T) \quad (81)$$

Therefore,

$$\|z\|_{1,\Omega}^2 \leq \sum_{T \in \mathcal{T}_h} \hat{a}_T(w_T, w_T)$$

Combining this with (77) yields the assertion. \square

3 FAST SOLVERS FOR THE hp-VERSION OF FE/BE COUPLING

In this section, we consider preconditioning techniques for the symmetric FE and BE coupling and apply the MINRES as the iterative solver; see Heuer, Maischak and Stephan (1999), Wathen and Stephan (1998) and see Mund and Stephan (1997), where its stable formulation, the HMCN is considered. For interface problems with second-order differential operators, this symmetric coupling procedure leads, as described above, to symmetric, indefinite linear systems (26), in which the condition numbers grow at least like $O(h^{-2}p^2)$ (see Maître and Poinquier, 1996). In Heuer and Stephan (1998) and Mund and Stephan (1998), the GMRES is analyzed, which applies to nonsymmetric problems but needs more storage and is therefore less efficient. A rather general preconditioner for BEM equations is formulated in Steinbach and Wendland (1997) (see also Steinbach and Wendland, 1998).

For brevity, we consider the interface problem in \mathbb{R}^2 for a polygonal domain Ω_1 and its complement $\Omega_2 = \mathbb{R}^2 \setminus \Omega_1$:

$$\begin{aligned} -\Delta u_1 &= f \quad \text{in } \Omega_1 \\ -\Delta u_2 &= 0 \quad \text{in } \Omega_2 \\ u_1 &= u_2 + u_0 \quad \text{on } \Gamma = \partial\Omega_1 \\ \frac{\partial u_1}{\partial n} &= \frac{\partial u_2}{\partial n} + t_0 \quad \text{on } \Gamma \end{aligned} \quad (82)$$

subject to the decay condition

$$u_2(x) = b \log |x| + o(1) \quad \text{as } |x| \rightarrow \infty \quad (83)$$

for a constant b . Here, $f \in H^{-1}(\Omega_1)$, $u_0 \in H^{1/2}(\Gamma)$, and $t_0 \in H^{-1/2}(\Gamma)$ are given functions. We assume that the interface Γ has conformal radius different from 1 such that the single-layer potential is injective. This can be achieved by an appropriate scaling of Ω_1 .

The proposed iterative method also applies to three-dimensional problems. Even an extension to nonlinear problems (see Carstensen and Stephan, 1995b) is straightforward because the use of the Newton-Raphson iteration reduces the task to solving a system of linear equations at each Newton step, which can be done by the same strategy, as discussed in this section. Moreover, the MINRES algorithm works for problems of linear and nonlinear elasticity.

In Heuer and Stephan (1998), we consider, instead of the Laplace equation in Ω_1 , the equation

$$-\Delta u_1 + k^2 u_1 = f \quad \text{in } \Omega_1$$

with $k > 0$, which yields a positive definite discretization of the corresponding Neumann problem. This was used in Heuer and Stephan (1998) to analyze the convergence of the GMRES method. In contrast, here we present the results from Heuer, Maischak and Stephan (1999) and show the convergence of the minimum residual method for an interface problem with positive semidefinite operator in Ω_1 (of course, the results of this section also hold for the operator $-\Delta + k^2$).

We approximate the solution of the interface problem by the hp-version with quasiniform meshes of the coupled FEM and BEM. This is based on the variational formulation (21), which is equivalent to (82) and (83) and is uniquely solvable because of Theorem 1 and Theorem 4.

Let us classify the basis functions that will be used in the Galerkin scheme (22). To this end, we introduce a uniform mesh of rectangles in Ω ,

$$\tilde{\Omega}_h = \bigcup_{j=1}^{J_0} \tilde{\Omega}_j$$

and use on Γ the mesh that is given by the restriction of $\tilde{\Omega}_h$ onto Γ ,

$$\tilde{\Gamma}_h = \bigcup_{j=1}^{J_r} \tilde{\Gamma}_j$$

However, we note that the two meshes $\tilde{\Omega}_h$ and $\tilde{\Gamma}_h$ need not be related, that is, $\tilde{\Omega}_h$ and $\tilde{\Gamma}_h$ can be chosen independently.

For our model problem, we consider polygonal domains that can be discretized by rectangular meshes. However, we note that triangular meshes can also be handled similarly by applying the decompositions proposed in Babuška *et al.* (1991). We note that the restriction to rectangular FE meshes does not influence the BE mesh, which is in either case the union of straight line pieces.

Let $\{S_j; j = 1, \dots, J_{\text{edges}}\}$ denote the set of edges of the mesh $\tilde{\Omega}_h$. We assume that the basis functions can be divided into the following four sets:

The set X_1 of the nodal functions. For each node of the mesh Ω_h , there is a function that has the value 1 at the node and is zero at the remaining nodes.

The sets X_S of the edge functions. For each edge S_j of the mesh Ω_h , there are functions vanishing at all other edges and which are nonzero only at the elements adjacent to S_j (and on S_j).

The sets X_{D_i} of the interior functions. For each element Ω_i of the mesh Ω_h , there are functions being nonzero only in the interior of Ω_i .

The sets of BE functions. For each element Γ_j of the BE mesh Γ_h , there are functions whose supports are contained in Γ_j . Note that the BE functions need not be continuous since $\gamma_N \in H^{-1/2}(\Gamma)$.

3.1 Preconditioned minimum residual method

In this section, we are concerned with the iterative solution of the linear system (26), in short form $Ax = b$. The coefficient matrix A is symmetric and indefinite. It is a discrete representation of a saddle-point problem. As an iterative solver, we use the preconditioned minimum residual method (see Ashby, Mantuffel and Saylor, 1990; Wathen, Fischer and Silvester, 1995).

The minimum residual method belongs to the family of Krylov subspace methods. Stable formulations of this method for symmetric and indefinite systems are given in Paige and Saunders (1975) and Chandra, Eisenstat and Schultz (1977). If we denote the iterates by x_k , $k = 0, 1, \dots$, there holds

$$\|b - Ax_k\|_1 = \min_{x \in \mathcal{K}_k(A, r_0)} \|b - Ax\|_1$$

where $\mathcal{K}_k(A, r_0) = \text{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\}$ denotes the Krylov subspace and r_0 is the initial residual $r_0 = b - Ax_0$. For a symmetric, positive definite preconditioner M , this relation becomes

$$\|b - Ax_k\|_{M^{-1}} = \min_{x \in \mathcal{K}_k(M^{-1}A, M^{-1}r_0)} \|b - Ax\|_{M^{-1}}$$

where $\|z\|_{M^{-1}}^2 = z^T M^{-1} z$ (see Wathen and Silvester, 1993; Wathen, Fischer and Silvester, 1995).

Owing to Lebedev (1969), for the residuals r_k , there holds

$$\left(\frac{\|r_k\|_1}{\|r_0\|_1} \right)^{1/k} \leq 2^{1/2k} \left(\frac{1 - \sqrt{bc/ad}}{1 + \sqrt{bc/ad}} \right)^{1/2} \quad (84)$$

when the set of eigenvalues E of $M^{-1}A$ is of the form $E \subset [-a, -b] \cup [c, d]$ with $-a < -b < 0 < c < d$ and $b - a = d - c$.

Therefore, the numbers of iterations of the preconditioned minimum residual method, which are required to solve (26) up to a given accuracy, are bounded by $O(\sqrt{bc/ad})^{-1}$.

In Wathen, Fischer and Silvester (1995), it is shown that the above result also holds for nonsymmetric inclusion sets of the form

$$[-a, -bN^{-\alpha}] \cup [cN^{-2\alpha}, d]$$

yielding

$$\left(\frac{\|r_k\|_1}{\|r_0\|_1} \right)^{1/k} \leq 1 - N^{-3\alpha/2} \sqrt{\frac{bc}{ad}}$$

3.2 Preconditioners

In the following section, we present preconditioners for (26) and give inclusion sets E for the eigenvalues of the preconditioned system matrices; see Heuer, Maischak and Stephan (1999) for the proofs of Theorem 8, Theorem 9, Theorem 10, and Theorem 11. The various preconditioners are 3-block and 2-block preconditioners or additive Schwarz preconditioners based on exact inversion of subblocks or on partially diagonal scaling. Either these preconditioned system matrices have bounded eigenvalues (for the 2-block method), in which the bounds do not depend on h or p or $E \subset [-a, -b] \cup [c, d]$ (for the additive Schwarz method and the partially diagonal scaling), where a, d are independent of h and p , whereas b, c are independent of h but behave like $O(p^{-\alpha}(1 + \log p)^{-\beta})$. By (84), the numbers of iterations of the minimum residual method remain bounded in the first situation, whereas the iteration numbers increase like $O(p^{\beta}(1 + \log p)^{\beta})$ in the latter cases.

The Galerkin matrix (26) belonging to the symmetric coupling method consists of a FE block, which is the discretization of an interior Neumann problem and of a BE block having the discretized hypersingular and weakly singular operators as terms on the diagonal.

For the analysis of preconditioners, one requires spectral estimates for the individual submatrices in (26). Denoting by $\Lambda(Q)$ the eigenvalue spectrum of a square matrix Q , there holds with suitable constants c_1, \dots, c_8 ,

$$\begin{aligned} \Lambda(A) &\subset [c_1 h^2 p^{-4}, c_2] \\ \Lambda(C) &\subset [c_3, c_4], \quad c_3 \geq 0 \\ \Lambda(W) &\subset [0, c_5] \\ \Lambda(V) &\subset [c_7 h^2 p^{-2}, c_8 h] \end{aligned}$$

Considering separately the FE functions on the interface boundary and in the interior domain and taking the BE

functions that discretize the weakly singular operator, we have a splitting of the ansatz space into subspaces. They induce a 3-block decomposition of the Galerkin matrix, which will be the 3-block preconditioner. By this decomposition, the strong coupling of edge and interior functions is neglected. Therefore, this 3-block splitting allows only for suboptimal preconditioners. Already in case of exactly inverting the blocks, one gets $O(h^{-3/4} p^{3/2})$ iteration numbers.

In detail, we employ a preconditioner of the form

$$M_3 = \begin{pmatrix} \tilde{A} & 0 & 0 \\ 0 & \tilde{C} & 0 \\ 0 & 0 & \tilde{V} \end{pmatrix} \quad (85)$$

where \tilde{A} , \tilde{C} , and \tilde{V} are spectrally equivalent matrices to A , C , and V respectively, that is, with suitable constants c_9, \dots, c_{14} , there holds

$$\begin{aligned} \Lambda(\tilde{A}^{-1/2} A \tilde{A}^{-1/2}) &\subset [c_9, c_{10}] \\ \Lambda(\tilde{C}^{-1/2} C \tilde{C}^{-1/2}) &\subset [c_{11}, c_{12}] \\ \Lambda(\tilde{V}^{-1/2} V \tilde{V}^{-1/2}) &\subset [c_{13}, c_{14}] \end{aligned}$$

For the symmetrically preconditioned matrix, denoted by $\tilde{A}_3 = M_3^{-1/2} A M_3^{-1/2}$, we have the following result.

Theorem 8. There exist positive constants a, b, c , and d , which are independent of h and p such that there holds

$$\Lambda(\tilde{A}_3) \subset [-a, -b] \cup [chp^{-2}, d]$$

Furthermore, the iteration numbers for the 3-block preconditioned minimum residual method grow like $O(h^{-3/4} p^{3/2})$.

Considering the Neumann block as a whole, that is, by taking together finite element functions on the interface and in the interior, we obtain a 2-block Jacobi method, which has bounded iteration numbers for exact inversion of the two blocks and therefore allows for almost optimal 2-block preconditioners.

In order to introduce the 2-block preconditioner, we use the following 2×2 -block representation for A :

$$A = \begin{pmatrix} A_N + W & \tilde{K}^T \\ \tilde{K} & -V \end{pmatrix} \quad \text{where} \quad A_N = \begin{pmatrix} A & B^T \\ B & C \end{pmatrix} \quad \tilde{K}^T = \begin{pmatrix} 0 \\ K^T - I \end{pmatrix}$$

A_N is a FE discretization of the homogeneous Neumann problem for the Laplacian (the subscript N in A_N refers to Neumann).

Our preconditioning matrix is

$$M_2 = \begin{pmatrix} \tilde{A}_M & 0 \\ 0 & \tilde{V} \end{pmatrix} \quad (86)$$

where \tilde{A}_M is spectrally equivalent to $A_N + W + M$ and \tilde{V} is spectrally equivalent to V . Here, M is an additional mass matrix, which is added to make $A_N + W$ positive definite. Then, the preconditioned matrix in 2×2 -block form is

$$\tilde{A}_2 = M_2^{-1/2} A M_2^{-1/2} = \begin{pmatrix} \tilde{A} & \tilde{K}^T \\ \tilde{K} & -\tilde{V} \end{pmatrix} \quad (87)$$

with

$$\tilde{A} = \tilde{A}_M^{-1/2} (A_N + W) \tilde{A}_M^{-1/2}, \quad \tilde{V} = \tilde{V}^{-1/2} V \tilde{V}^{-1/2} \\ \tilde{K}^T = \tilde{A}_M^{-1/2} K^T \tilde{V}^{-1/2}$$

Theorem 9. There exist positive constants a, b, c , and d , which are independent of h and p such that there holds

$$\Lambda(\tilde{A}_2) \subset [-a, -b] \cup [c, d]$$

Furthermore, the number of iterations of the 2-block preconditioned minimum residual method is bounded independently of h and p .

Remark 2. For $\tilde{V} = V$, all eigenvalues λ_i are 1, yielding $b = 1$ in Theorem 9.

The additive Schwarz preconditioner extends the 2-block method by replacing the main blocks by block diagonal matrices. Here we proceed as follows. First, we construct discrete harmonic functions by applying the Schur complement method for the FE block of the Galerkin matrix. Then, for the FE part, we decompose the test and trial functions in nodal, edge, and interior functions. This amounts for the finite element block to a block Jacobi (Additive Schwarz) preconditioner. We split the BE block belonging to the weakly singular integral operator with respect to unknowns belonging to a coarse grid space consisting of piecewise constant functions, and belonging to individual subspaces according to the BE, consisting of all polynomials up to degree p without the constants. Our second preconditioner is obtained by further splitting the subspaces of edge functions (for both the FE and the BE) into one-dimensional subspaces according to the edge-basis functions. For the 2-block method, we obtain in this way two different preconditioned linear systems, which need respectively $O(\log^2 p)$ and $O(p \log^2 p)$ minimum residual iterations to be solved up to a given accuracy. We note that we are dealing with direct sum decompositions only. Having ensured the edge functions to be discrete harmonic, all

the arising local problems are independent of each other. Therefore, the procedures are capable of being parallelized without special modifications. Also, the Schur complement procedure can be parallelized on the elements' level (cf. Babuška *et al.*, 1991).

In the notation of (86), we take matrices \tilde{A}_M and \tilde{V} , which are spectrally equivalent to $A_N + W + M$ and V respectively. However, the respective equivalence constants will depend on p , but not on h . Since we define the blocks by decomposing the subspaces X_M and Y_N , this method is referred to as an additive Schwarz method. The decomposition of the FE space X_N is given in Babuška *et al.* (1991) and the decomposition of the BE space has been proposed in Tran and Stephan (2000). This decomposition of the ansatz spaces into subspaces of nodal functions, edge functions for each edge, interior functions for each FE, and into subspaces for each BE is as follows:

$$X_M = X_1 \cup X_{S_1} \cup \dots \cup X_{S_{j_{\text{edge}}}} \cup X_{\Omega_1} \cup \dots \cup X_{\Omega_M} \quad (88)$$

and

$$Y_N = Y_0 \cup Y_{\Gamma_1} \cup \dots \cup Y_{\Gamma_{j_N}} \quad (89)$$

Here, the space X_1 is the space of the nodal functions, X_{S_j} is the space of the edge functions related to the edge S_j , and X_{Ω_j} is spanned by the interior functions on the element Ω_j . For the BE functions, we assume that Y_0 consists of the piecewise constant functions on the mesh Γ_h , and Y_{Γ_j} is spanned by the Legendre polynomials l_i , $i = 1, \dots, p-1$, mapped onto the element Γ_j .

The preconditioner that belongs to the decompositions (88) and (89) is denoted by M_{ASM} . It is the block diagonal matrix of A , in which the blocks belonging to the subspaces in the decompositions are taken. In accordance to the 2-block method, we denote the finite element and BE parts of the additive Schwarz preconditioner M_{ASM} by $A_M = A_{ASM}$ and $\tilde{V} = V_{ASM}$.

The following theorem shows the impact of the additive Schwarz preconditioner on the spectrum of the coupling matrix and together with (84) leads to a growth for the iteration numbers like $O(\log^2 p)$.

Theorem 10. (i) There holds

$$\Lambda(A_{ASM}^{-1}(A_N + M)) \cup \Lambda(V_{ASM}^{-1}V) \subset [c(1 + \log p)^{-2}, C]$$

for constants $c, C > 0$, which are independent of h and p .

(ii) Let the set of nodal functions of X_M be spanned by the standard piecewise bilinear functions and assume the edge functions to be discrete harmonic, that is, for $i = 1, \dots, j_{\text{edge}}$, and $j = 1, \dots, j_M$

$$a(u, v) = 0 \quad \text{for all } u \in X_{S_j}, v \in X_{\Omega_j}$$

where $a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v$. Then, there exist positive constants a, b, c , and d , which are independent of h and p such that there holds

$$\begin{aligned} \Lambda(M_{ASM}^{-1}A) &\subset [-a, -b(1 + \log p)^{-2}] \\ &\cup [c(1 + \log p)^{-2}, d] \end{aligned}$$

A preconditioner based on *partially diagonal scaling* is obtained by further refining the subspace decompositions. We take the block diagonal preconditioner, which consists of the blocks belonging separately to the piecewise bilinear functions, the interior functions for individual elements, and the piecewise constant functions. For the remaining functions, we simply take the diagonal of the stiffness matrix A . This method combines the decompositions of X_M proposed in Babuška *et al.* (1991) and of Y_N proposed in Heuer, Stephan and Tran (1998). More precisely, we take the decompositions

$$X_M = X_1 \cup \bigcup_{j=1}^{j_{\text{edge}}} \bigcup_{q=2}^p X_{S_{j,q}} \cup X_{\Omega_1} \cup \dots \cup X_{\Omega_M} \quad (90)$$

and

$$Y_N = Y_0 \cup \bigcup_{j=1}^{j_N} \bigcup_{q=1}^{p-1} Y_{\Gamma_{j,q}} \quad (91)$$

The subspaces X_1 , X_{Ω_j} ($j = 1, \dots, j_M$), and Y_0 are as before, whereas $X_{S_{j,q}}$ ($j = 1, \dots, j_{\text{edge}}$) and $Y_{\Gamma_{j,q}}$ ($j = 1, \dots, j_N$) consist of individual edge functions (degree ≥ 2) and BE functions (degree ≥ 1) respectively.

In accordance with M_{ASM} , we define the preconditioner M_{diag} , which is the block diagonal matrix consisting of the blocks of A belonging to the subspaces in (90) and (91). In the notation of the 2-block method, we have $\tilde{A}_M = A_{\text{diag}}$ and $\tilde{V} = V_{\text{diag}}$.

Theorem 11. (i) There holds

$$\Lambda(A_{\text{diag}}^{-1}(A_N + M)) \subset [cp^{-1}(1 + \log p)^{-3}, C]$$

and

$$\Lambda(V_{\text{diag}}^{-1}V) \subset [cp^{-1}(1 + \log p)^{-2}, C]$$

for constants $c, C > 0$, which are independent of h and p . (ii) Let the set of nodal functions of X_M be spanned by the standard piecewise bilinear functions and assume the edge functions to be discrete harmonic.

Then, there exist positive constants a, b, c , and d , which are independent of h and p such that there holds

$$\begin{aligned} \Lambda(M_{\text{diag}}^{-1}A) &\subset [-a, -bp^{-1}(1 + \log p)^{-2}] \\ &\cup [cp^{-1}(1 + \log p)^{-2}, d] \end{aligned}$$

With this preconditioner, we obtain for the iteration numbers of the minimum residual method a growth like $O(p \log^2 p)$.

3.3 Implementation issues and numerical results

For the spaces Y_N , we use discontinuous piecewise Legendre polynomials on the decomposition of Γ , and for the spaces X_M , we use tensor products of antiderivatives of the Legendre polynomials on each of the rectangles. The antiderivatives of the Legendre polynomials with a degree of at least 1 are automatically globally continuous as they vanish at the endpoints of their supports. The piecewise linear FE functions have to be assembled such that they are continuous on Ω . To assemble a finite-dimensional subspace $X_M \times Y_N$ for the coupling method, we take a degree p and construct X_M by assembling for all rectangles the tensor products $f_i \times f_j$ of polynomials of degrees i and j respectively, up to $\max(i, j) = p$. For Y_N , we use on all BE all the Legendre polynomials up to degree $p-1$. This discretization yields a linear system of the form (26). Here, u_M are the unknowns with respect to the functions of X_M interior to Ω , u_M are the unknowns with respect to the functions of X_M on the boundary, and ϕ_N represents the unknowns belonging to Y_N .

The implementation of the A, B, C, I -blocks is done analytically. The bilinear forms involving integral operators, that is, represented by the blocks W, V, K, K^T , are rewritten in such a way that the inner integration can be carried out again analytically. The outer integration is performed numerically by a Gaussian quadrature. For more details, see Ervin, Heuer and Stephan (1993) and Maischak, Stephan and Tran (2000).

We solve the linear system (26) via the minimum residual method, in which we consider the un-preconditioned version and the preconditioners M_3, M_2, M_{ASM} , and M_{diag} . The theoretical results for the preconditioners M_{ASM} and M_{diag} require discrete harmonic edge functions, that is, they have to be orthogonal with respect to the H^1 -inner product to the basis functions, which are interior to the elements. This is fulfilled by performing a Schur complement step with respect to the interior basis functions, resulting in a basis transformation of the edge functions. For performing the Schur complement step, the inversions of the blocks of the interior functions are done directly, as it is done for all the blocks of the preconditioners. For practical applications, these direct inversions may be replaced by indirect solvers. The action of performing the Schur complements is a local operation, which can be parallelized on the elements' level and is therefore not a very time-consuming task. On the other hand, it is also possible to choose edge-basis functions, which are a priori discrete harmonic (see Heuer and Stephan, 2001).

In Heuer, Maischak and Stephan (1999), numerical experiments are presented for the interface problem (82) and (83) with the L-shaped domain Ω_1 with vertices $(0, 0)$, $(0, 1/2)$, $(-1/2, 1/2)$, $(-1/2, -1/2)$, $(1/2, -1/2)$, and $(1/2, 0)$ and given data u_0 and f_0 chosen such that

$$u_1(x, y) = \Im(z^{2/3}) \quad \text{for } z = x + iy$$

and

$$u_2(x, y) = \log |(x, y) + (0.3, 0.3)|$$

The numerical experiments listed in the paper by Heuer, Maischak and Stephan (1999) underline the above results on the spectral behavior of the Galerkin matrix and the various preconditioned versions. Here, we just list in Table 2 (from Heuer, Maischak and Stephan, 1999) the numbers of iterations of the minimum residual method that are required to reduce the initial residual by a factor of 10^{-3} . One observes that for the un-preconditioned linear system, the numbers of iterations increase rather fast, making the system almost intractable for large dimensions of the ansatz spaces X_M and Y_N . On the other hand, the 2-block preconditioner keeps the numbers of iterations bounded at a rather moderate value. The 3-block preconditioner, which is almost as expensive as the 2-block method (since in both cases at least the block belonging to the interior FE functions needs to be inverted), results in increasing iteration numbers that are comparable with the numbers necessary for the additive Schwarz preconditioner. Here, in all cases, the various sub-blocks are inverted exactly. As expected, the numbers of iterations that are necessary for the partially diagonal scaling are larger than all of the other preconditioners. On the other hand, this method is the cheapest one and, in comparison with the un-preconditioned method, it reduces the numbers of iterations substantially.

Table 2. Numbers of MINRES iterations required to reduce the initial residual by a factor of 10^{-3} .

| $1/h$ | p | $M+N$ | A | $M_3^{-1}A$ | $M_2^{-1}A$ | $M_{ASM}^{-1}A$ | $M_{\text{diag}}^{-1}A$ |
|-------|-----|-------|-------|-------------|-------------|-----------------|-------------------------|
| 2 | 2 | 37 | 36 | 21 | 11 | 30 | 30 |
| 2 | 4 | 97 | 138 | 34 | 12 | 47 | 52 |
| 2 | 6 | 181 | 285 | 40 | 13 | 54 | 67 |
| 2 | 8 | 289 | 536 | 45 | 13 | 62 | 77 |
| 2 | 10 | 421 | 892 | 50 | 13 | 66 | 94 |
| 2 | 12 | 577 | 1374 | 55 | 13 | 74 | 111 |
| 2 | 14 | 757 | 2328 | 60 | 13 | 82 | 135 |
| 2 | 16 | 961 | >9999 | 79 | 17 | 102 | 197 |
| 4 | 1 | 37 | 17 | 15 | 10 | | |
| 8 | 1 | 97 | 38 | 23 | 13 | | |
| 16 | 1 | 289 | 67 | 32 | 13 | | |
| 32 | 1 | 961 | 130 | 45 | 14 | | |
| 64 | 1 | 3457 | 243 | 60 | 15 | | |
| 128 | 1 | 13057 | 476 | 75 | 15 | | |

4 LEAST SQUARES FE/BE COUPLING METHOD

We introduce a least squares FE/BE coupling formulation for the numerical solution of second-order linear transmission problems in two and three dimensions, which allow jumps on the interface, (see Maischak and Stephan). In a bounded domain, the second-order partial differential equation is rewritten as a first-order system; the part of the transmission problem that corresponds to the unbounded exterior domain is reformulated by means of boundary integral equations on the interface. The least squares functional is given in terms of Sobolev norms of the orders -1 and $1/2$. In case of the h -version of the FE/BE coupling, these norms are computed by using multilevel preconditioners for a second-order elliptic problem in a bounded domain Ω , and for the weakly singular integral operator of the single-layer potential on its boundary $\partial\Omega$. We use both MG and BPX algorithms as preconditioners, and the preconditioned system has bounded or mildly growing condition number. These preconditioners can be used to accelerate the computation of the solution of the full discrete least squares system by a preconditioned conjugate gradient method. Thus, this least squares coupling approach gives a fast and robust solution procedure. The given approach should be applicable to more general interface problems from elasticity and electromagnetics. Numerical experiments confirm our theoretical results (cf. Table 3 and Figure 2).

Here we consider again the interface problem (IP), that is, (1) to (5), for $A(\nabla u_1) = a \nabla u_1$, where $a_{ij} \in L^\infty(\Omega)$ such that there exists a $\alpha > 0$ with

$$\alpha \|z\|^2 \leq z^T a(x) z \quad \forall z \in \mathbb{R}^d \text{ and for almost all } x \in \Omega$$

Introducing the flux variable $\theta := a \nabla u_1$ and the new unknown $\sigma := (a \nabla u_1) \cdot n$, we note that the unknown θ belongs to $H(\operatorname{div}; \Omega)$, where

$$H(\operatorname{div}; \Omega) = \{\theta \in [L^2(\Omega)]^d : \|\theta\|_{L^2(\Omega)}^2 + \|\operatorname{div} \theta\|_{L^2(\Omega)}^2 < \infty\}$$

With the inner product

$$(\theta, \xi)_{H(\operatorname{div}; \Omega)} = (\theta, \xi)_{L^2(\Omega)} + (\operatorname{div} \theta, \operatorname{div} \xi)_{L^2(\Omega)}$$

$H(\operatorname{div}; \Omega)$ is a Hilbert space. Moreover, for all $\xi \in H(\operatorname{div}; \Omega)$, there holds $\xi \cdot n \in H^{-1/2}(\Gamma)$ and $\|\xi \cdot n\|_{H^{-1/2}(\Gamma)} \leq \|\xi\|_{H(\operatorname{div}; \Omega)}$ (see Girault and Raviart, 1986).

Incorporating the interface conditions, we can rewrite the transmission problem (IP), (1) to (5), into the following formulation with first-order system on Ω .

Find $(\theta, u, \sigma) \in H(\operatorname{div}; \Omega) \times H^1(\Omega) \times H^{-1/2}(\Gamma)$ such that

$$\theta = a \nabla u \quad \text{in } \Omega \quad (92)$$

$$-\operatorname{div} \theta = f \quad \text{in } \Omega \quad (93)$$

$$\sigma = \theta \cdot n \quad \text{on } \Gamma \quad (94)$$

$$2(\sigma - t_0) = -W(u - u_0) + (I - K')(\sigma - t_0) \quad \text{on } \Gamma \quad (95)$$

$$0 = (I - K')(u - u_0) + V(\sigma - t_0) \quad \text{on } \Gamma \quad (96)$$

In the following, let $\tilde{H}^{-1}(\Omega)$ denote the dual space of $H^1(\Omega)$, equipped with the norm $\|w\|_{\tilde{H}^{-1}(\Omega)} = \sup_{v \in H^1(\Omega)} ((w, v)_{L^2(\Omega)}) / \|v\|_{H^1(\Omega)}$.

We observe that the solution of (92) to (96) is a solution of the following quadratic minimization problem.

Find $(\theta, u, \sigma) \in X := [L^2(\Omega)]^d \times H^1(\Omega) \times H^{-1/2}(\Gamma)$ such that

$$J(\theta, u, \sigma) = \min_{(\xi, v, \tau) \in X} J(\xi, v, \tau) \quad (97)$$

where J is the quadratic functional defined by

$$\begin{aligned} J(\xi, v, \tau) &= \|a \nabla v - \xi\|_{L^2(\Omega)}^2 \\ &+ \|(I - K')(v - u_0) + V(\tau - t_0)\|_{H^{1/2}(\Gamma)}^2 \\ &+ \|\operatorname{div} \xi + f - \frac{1}{2} \delta_\Gamma \otimes (W(v - u_0) \\ &+ 2\xi \cdot n - 2t_0 - (I - K')(\tau - t_0))\|_{\tilde{H}^{-1}(\Omega)}^2 \\ &= \|a \nabla v - \xi\|_{L^2(\Omega)}^2 + \|(I - K')v \\ &+ V\tau - (I - K')u_0 - Vt_0\|_{H^{1/2}(\Gamma)}^2 \\ &+ \|\operatorname{div} \xi - \frac{1}{2} \delta_\Gamma \otimes (Wv + 2\xi \cdot n - (I - K')\tau) \\ &+ f + \frac{1}{2} \delta_\Gamma \otimes (Wu_0 + 2t_0 - (I - K')t_0)\|_{\tilde{H}^{-1}(\Omega)}^2 \end{aligned} \quad (98)$$

Here, $\delta_\Gamma \otimes \tau$ denotes a distribution in $\tilde{H}^{-1}(\Omega)$ for $\tau \in H^{-1/2}(\Gamma)$. By proving coercivity and continuity of the corresponding variational problem, we obtain uniqueness of (97), and therefore we have the equivalence of (92) to (96) and (97).

Defining $g(\xi, v, \tau) := \operatorname{div} \xi - (1/2) \delta_\Gamma \otimes (Wv + 2\xi \cdot n - (I - K')\tau)$, we can write for the bilinear form corresponding to $J(\xi, v, \tau)$

$$\begin{aligned} B((\theta, u, \sigma), (\xi, v, \tau)) &= (a \nabla u - \theta, a \nabla v - \xi)_{L^2(\Omega)} \\ &+ ((I - K')u + V\sigma, (I - K')v + V\tau)_{H^{1/2}(\Gamma)} \\ &+ (g(\theta, u, \sigma), g(\xi, v, \tau))_{\tilde{H}^{-1}(\Omega)} \end{aligned} \quad (99)$$

and the linear functional

$$\begin{aligned} G(\xi, v, \tau) &= ((I - K')v + V\tau, (I - K')u_0 + Vt_0)_{H^{1/2}(\Gamma)} \\ &- (g(\xi, v, \tau), f + \frac{1}{2} \delta_\Gamma \\ &\otimes (Wu_0 + 2t_0 - (I - K')t_0))_{\tilde{H}^{-1}(\Omega)} \end{aligned} \quad (100)$$

The variational formulation now reads as follows:

Find $(\theta, u, \sigma) \in X = [L^2(\Omega)]^d \times H^1(\Omega) \times H^{-1/2}(\Gamma)$ such that

$$B((\theta, u, \sigma), (\xi, v, \tau)) = G(\xi, v, \tau) \quad \forall (\xi, v, \tau) \in X \quad (101)$$

In Maischak and Stephan, we prove the following result.

Theorem 12. *The bilinear form $B(\cdot, \cdot)$ is continuous and strongly coercive in $X \times X$ and the linear form $G(\cdot)$ is continuous on X . There exists a unique solution of the variational least squares formulation (101), which is also a solution of (92) to (96).*

Choosing subspaces $V_h \subset H^1(\Omega)$, $S_h \subset H^{-1/2}(\Gamma)$, $H_h \subset [L^2(\Omega)]^d$, one can pose a discrete formulation corresponding to (101) with a discrete bilinear form $B^{(h)}(\cdot, \cdot)$, which is again uniquely solvable.

For its implementation, we have to choose basis functions $\{\phi_i\}$ of V_h , basis functions $\{\lambda_i\}$ of S_h , and basis functions $\{\psi_i\}$ of H_h . In case of a h -version, later on we choose hat-functions for the discretization of V_h , piecewise constant functions, that is, brick-functions, for S_h , and investigate the use of hat-functions, brick-functions, and Raviart-Thomas (RT) elements for the discretization of the flux space H_h .

The introduction of basis functions leads to the definition of the following matrices and vectors.

$$\begin{aligned} (A_h)_{ij} &= (a \nabla \phi_i, a \nabla \phi_j)_{L^2(\Omega)}, & (F_h)_{ij} &= (a \nabla \phi_i, \psi_j)_{L^2(\Omega)} \\ (G_h)_{ij} &= (\psi_i, \psi_j)_{L^2(\Omega)}, & (f_h)_{ij} &= (f, \psi_j)_{L^2(\Omega)} \end{aligned}$$

For the BE part, we need the well-known dense matrices

$$\begin{aligned} (V_h)_{ij} &= \langle \lambda_i, V \lambda_j \rangle, & (K_h)_{ij} &= \langle \lambda_i, K \phi_j \rangle \\ (W_h)_{ij} &= \langle \phi_i, W \phi_j \rangle, & (I_h)_{ij} &= \langle \phi_i, \lambda_j \rangle \end{aligned}$$

The matrix representation of the discrete bilinear form $B^{(h)}(\cdot, \cdot)$ and the discrete linear form then becomes

$$\begin{aligned} \begin{bmatrix} G_h & -F_h^T & 0 \\ -F_h & A_h & 0 \\ 0 & 0 & 0 \end{bmatrix} &+ \begin{bmatrix} F_h^T \\ \frac{1}{2} W_h \\ \frac{1}{2} (K_h - I_h) \end{bmatrix} \\ &\times B_h [F_h, \frac{1}{2} W_h, \frac{1}{2} (K_h - I_h)^T] \end{aligned}$$

$$\begin{aligned} &+ \begin{bmatrix} 0 \\ (I_h - K_h)^T \\ V_h \end{bmatrix} C_h [0, I_h - K_h, V_h] \begin{bmatrix} \theta_h \\ u_h \\ \sigma_h \end{bmatrix} \\ &= \begin{bmatrix} F_h^T \\ \frac{1}{2} W_h \\ \frac{1}{2} (K_h - I_h) \end{bmatrix} B_h (f_h + \frac{1}{2} [(K' + I)u_0 + Wu_0]_h) \\ &- \begin{bmatrix} 0 \\ (I_h - K_h)^T \\ V_h \end{bmatrix} C_h [(K - I)u_0 - Vt_0]_h \end{aligned} \quad (102)$$

where $[\cdot]_h$ denotes testing with the bases functions. Here, B_h and C_h are preconditioners for the matrices $A_h + M_h$ and V_h , where M_h is the mass matrix.

We prove in Maischak and Stephan the following theorem dealing with the preconditioning for the discrete system (102). In applications, the matrix E_h should be an easily computable approximation of the inverse of the mass matrix, for example, a scaled identity matrix.

Theorem 13. *Let E_h be such that $(E_h^{-1} \xi_h, \xi_h)_{L^2(\Omega)} \sim (\xi_h, \xi_h)_{L^2(\Omega)}$. Then, with the preconditioners B_h and C_h , there holds the equivalence*

$$\begin{aligned} (E_h^{-1} \xi_h, \xi_h)_{L^2(\Omega)} + (B_h^{-1} u_h, u_h)_{L^2(\Omega)} + (C_h^{-1} \tau_h, \tau_h)_{L^2(\Gamma)} \\ \sim B^{(h)}((\xi_h, u_h, \tau_h), (\xi_h, u_h, \tau_h)) \end{aligned}$$

Therefore, $\operatorname{diag}(E_h^{-1}, B_h^{-1}, C_h^{-1})$ is spectrally equivalent to the system matrix $B^{(h)}(\cdot, \cdot)$, and if the block diagonal matrix $\operatorname{diag}(E_h, B_h, C_h)$ is applied to the system (102), the resulting system has a bounded condition number.

B_h denotes the preconditioner for the FE matrix A_h stabilized by the mass matrix M_h , $(M_h)_{ij} = (\phi_i, \phi_j)_{L^2(\Omega)}$ (see Funken and Stephan, 2001; Heuer, Maischak and Stephan, 1999), and C_h is the preconditioner for the matrix with the single-layer potential V_h . For B_h and C_h , we use MG, BPX, and the inverse matrices $(A_h + M_h)^{-1}$ and V_h^{-1} (INV) (performed by several MG steps). The MG algorithm for B_h gives a preconditioner, which is spectrally equivalent to the inverse of the above stabilized FE matrix A_h , whereas BPX for B_h leads to a linear system with a mildly growing condition number. In case of C_h , we use, as in Bramble, Leyk and Pasciak (1992), the multilevel algorithm (MG and BPX), which incorporates the second-order difference operator. Another preconditioner C_h , which is spectrally equivalent to the inverse of the single-layer potential matrix V_h up to terms depending logarithmically on h , is given by using additive Schwarz method based on the Haar basis (Tran and Stephan, 1996; Maischak, Stephan and Tran, 1996). As the linear system solver, we take the preconditioned conjugate gradient algorithm until the relative change of the iterated solution is less than $\delta = 10^{-8}$.

The implementation of the least squares coupling method uses only components that are also necessary in the implementation of the standard symmetric coupling method and offers the advantage that there is no need for a generalized Krylov method like MINRES, as in Section 3, or GMRES, as in the case of the symmetric coupling method; see Funken and Stephan (2001) and Mund and Stephan (1998). The well-known preconditioned conjugate gradient algorithm is sufficient. The computation of the flux is done with negligible implementation effort because no preconditioner is needed (cf. Theorem 13), and the Galerkin matrices involved are sparse and very easy to implement. The numerical experiments presented below are done with the software package *maiprog*s (see Maischak, 2003). For further numerical experiments with piecewise constant and piecewise linear fluxes, see Maischak and Stephan.

To underline the above approach, we present least squares FE/BE coupling computations from Maischak and Stephan, in which, on a uniform triangulation with rectangles, θ is computed with $H(\text{div}; \Omega)$, conforming RT elements of the lowest order, u_h is piecewise bilinear and σ_h is piecewise constant.

In Table 3, the corresponding L^2 -errors $\delta_\theta, \delta_u, \delta_\sigma$ are given for the following example.

Let Ω be the L-shaped domain with vertices $(0, 0)$, $(0, 1/4)$, $(-1/4, 1/4)$, $(-1/4, -1/4)$, $(1/4, -1/4)$, and $(1/4, 0)$. Now, we prescribe jumps with singularities on the interface $\Gamma = \partial\Omega$ and take $f = 0$ in Ω and $a \equiv 1$. Setting

$$u_0(r, \phi) := r^{2/3} \sin \left[\frac{2}{3}(2\pi - \phi) \right] \\ - \log \sqrt{\left(x - \frac{1}{8}\right)^2 + \left(y - \frac{1}{8}\right)^2}, \quad t_0(r, \phi) := \frac{\partial u_0}{\partial n}$$

the solution of the interface problem (IP), that is, (1) to (5), is given by

$$u_1(r, \phi) = r^{2/3} \sin \left[\frac{2}{3}(2\pi - \phi) \right] \quad \text{in } \Omega_1$$

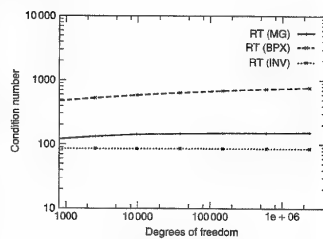


Figure 2. Condition numbers, θ_h , with RT elements.

$$u_2(r, \phi) = \log \sqrt{\left(x - \frac{1}{8}\right)^2 + \left(y - \frac{1}{8}\right)^2} \quad \text{in } \Omega_2$$

The experimental convergence rates given in Table 3 confirm the theoretical convergence rates. In Figure 2, the condition numbers for the preconditioned system (102) are plotted, showing excellent behavior for the block-inverse and the MG preconditioner, whereas BPX slowly degenerates.

5 FE/BE COUPLING FOR INTERFACE PROBLEMS WITH SIGNORINI CONTACT

5.1 Primal method

Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be a bounded domain with Lipschitz boundary Γ . Let $\Gamma = \Gamma_1 \cup \Gamma_2$, where Γ_1 and Γ_2 are nonempty, disjoint, and open in Γ . In the interior part, we consider a nonlinear partial differential equation, whereas

in the exterior part, we consider the Laplace equation

$$-\text{div}(q(|\nabla u|) \cdot \nabla u) = f \quad \text{in } \Omega \quad (103)$$

$$-\Delta u = 0 \quad \text{in } \Omega_c = \mathbb{R}^d \setminus \bar{\Omega} \quad (104)$$

with the radiation condition

$$u(x) = O(|x|^{2-d}) \quad \text{for } d \geq 2, \quad (|x| \rightarrow \infty) \quad (105)$$

Here, $q: [0, \infty) \rightarrow [0, \infty)$ is a $C^1[0, \infty)$ function with $t \cdot q(t)$ being monotonously increasing with t , $q(t) \leq 0$, $(t \cdot q(t))' \leq q_1$ and let $q(t) + t \cdot \min(0, q(t)) \geq \alpha > 0$.

Writing $u_1 := u|_{\Omega}$ and $u_2 := u|_{\Omega_c}$, the tractions on Γ are given by $q(|\nabla u_1|) \frac{\partial u_1}{\partial n}$ and $-(\frac{\partial u_2}{\partial n})$ with normal n pointing into Ω_c .

We consider transmission conditions on Γ_1 :

$$u_1|_{\Gamma_1} - u_2|_{\Gamma_1} = u_0|_{\Gamma_1}$$

and

$$q(|\nabla u_1|) \frac{\partial u_1}{\partial n} \Big|_{\Gamma_1} - \frac{\partial u_2}{\partial n} \Big|_{\Gamma_1} = t_0|_{\Gamma_1} \quad (106)$$

and Signorini conditions on Γ_2

$$u_1|_{\Gamma_2} - u_2|_{\Gamma_2} \leq u_0|_{\Gamma_2}$$

$$q(|\nabla u_1|) \frac{\partial u_1}{\partial n} \Big|_{\Gamma_2} = \frac{\partial u_2}{\partial n} \Big|_{\Gamma_2} + t_0|_{\Gamma_2} \leq 0 \quad (107)$$

$$0 = q(|\nabla u_1|) \frac{\partial u_1}{\partial n} \Big|_{\Gamma_2} \cdot (u_2 + u_0 - u_1)|_{\Gamma_2}$$

Given data $f \in L^2(\Omega)$, $u_0 \in H^{1/2}(\Gamma)$, and $t_0 \in H^{-1/2}(\Gamma)$ (with $(f, 1)_{L^2(\Omega)} + (t_0, 1) = 0$ if $d = 2$), we look for $u_1 \in H^1(\Omega)$ and $u_2 \in H_{\text{loc}}^1(\Omega_c)$, satisfying (103) to (107) in a weak form.

Setting

$$g(t) = \int_0^t s \cdot q(s) \, ds$$

the assumptions on q yield that

$$G(u) = 2 \int_{\Omega} g(|\nabla u|) \, dx$$

is finite for any $u \in H^1(\Omega)$ and its Fréchet derivative

$$DG(u; v) = 2 \int_{\Omega} q(|\nabla u|) (\nabla u)^T \cdot \nabla v \, dx \quad \forall u, v \in H^1(\Omega) \quad (108)$$

is uniformly monotone, that is, there exists a constant $\gamma > 0$ such that

$$\gamma \|u - v\|_{H^1(\Omega)}^2 \leq DG(u; u - v)$$

$$-DG(v; u - v) \quad \forall u, v \in H^1(\Omega) \quad (109)$$

(see Carstensen and Gwinner (1997, Proposition 2.1).

Following Carstensen and Gwinner (1997), Maischak (2001) analyses an FE/BE coupling procedure for (103) to (107), which uses the Steklov–Poincaré operator S (32) for the exterior problem.

Let $E := H^1(\Omega) \times H_{00}^{1/2}(\Gamma_1)$, where $H_{00}^{1/2}(\Gamma_1) := \{w \in H^{1/2}(\Gamma) : \text{supp } w \subseteq \Gamma_1\}$ and set

$$D := \{(u, v) \in E : v \geq 0 \text{ a.e. on } \Gamma_2 \\ \text{and } \langle S1, u|_{\Gamma_1} + v - u_0 \rangle = 0 \text{ if } d = 2\}$$

Then, the primal formulation of (103) to (107), called problem (SP), consists in finding (\hat{u}, \hat{v}) in D such that

$$\Psi(\hat{u}, \hat{v}) = \inf_{(u, v) \in D} \Psi(u, v)$$

where

$$\Psi(u, v) := 2 \int_{\Omega} g(|\nabla u|) \, dx \\ + \frac{1}{2} \langle S(u|_{\Gamma_1} + v), u|_{\Gamma_1} + v \rangle - \lambda(u, v)$$

and $\lambda \in E^*$, the dual of E , is given by

$$\lambda(u, v) := L(u, u|_{\Gamma_1} + v) + \langle Su_0, u|_{\Gamma_1} + v \rangle$$

with

$$L(u, v) := 2 \int_{\Omega} f \cdot u \, dx + 2 \int_{\Gamma} t_0 \cdot v \, ds$$

for any $(u, v) \in E$.

Owing to Carstensen and Gwinner (1997), there exists exactly one solution $(\hat{u}, \hat{v}) \in D$ of problem (SP), which is the variational solution of the transmission problem (103) to (107). Moreover, $(\hat{u}, \hat{v}) \in D$ is the unique solution of the variational inequality

$$A(\hat{u}, \hat{v})(u - \hat{u}, v - \hat{v}) \geq \lambda(u - \hat{u}, v - \hat{v}) \quad (110)$$

for all $(u, v) \in D$, with

$$A(u, v)(r, s) := DG(u, r) + \langle S(u|_{\Gamma_1} + v), r|_{\Gamma_1} + s \rangle \quad (111)$$

For the discretization, we take nested regular quasiuniform meshes $(\mathcal{T}_h)_h$ consisting of triangles or quadrilaterals. Then, let H_h^1 denote the related continuous and piecewise affine trial functions on the triangulation \mathcal{T}_h . The mesh on Ω induces a mesh on the boundary, so that we may consider $H_h^{1/2}$ as the piecewise constant trial functions. Assuming that the partition of the boundary also leads to a partition of Γ_1 , $H_h^{1/2}$ is then the subspace of continuous and piecewise linear functions on the partition of Γ_1 , which vanish at intersection points in

Table 3. L^2 -errors $\delta_\theta, \delta_u, \delta_\sigma$ for θ_h, u_h, σ_h and convergence rates (θ_h with RT, MG preconditioner).

| #total | h | δ_θ | α_θ | δ_u | α_u | δ_σ | α_σ |
|---------|---------|-----------------|-----------------|------------|------------|-----------------|-----------------|
| 209 | 0.06250 | 0.05517 | | 0.0008359 | | 0.45692 | |
| 705 | 0.03125 | 0.03535 | 0.642 | 0.0002838 | 1.558 | 0.40826 | 0.162 |
| 2561 | 0.01562 | 0.02249 | 0.653 | 0.9398E-04 | 1.594 | 0.36480 | 0.162 |
| 9729 | 0.00781 | 0.01425 | 0.658 | 0.3147E-04 | 1.578 | 0.32598 | 0.162 |
| 37889 | 0.00390 | 0.00901 | 0.661 | 0.1203E-04 | 1.387 | 0.29128 | 0.162 |
| 149505 | 0.00195 | 0.00569 | 0.663 | 0.6128E-05 | 0.973 | 0.26026 | 0.162 |
| 593921 | 0.00097 | 0.00359 | 0.665 | 0.3801E-05 | 0.689 | 0.23248 | 0.163 |
| 2367489 | 0.00048 | 0.00226 | 0.665 | 0.2468E-05 | 0.623 | 0.20758 | 0.163 |

$\tilde{\Gamma}_s \cap \tilde{\Gamma}_t$. Then, we have $H_h^1 \times \tilde{H}_h^{1/2} \times H_h^{-1/2} \subset H^1(\Omega) \times \tilde{H}_0^{1/2}(\Gamma_s) \times H^{-1/2}(\Gamma)$. Now, D_h is given by

$$D_h := \{(u_h, v_h) \in H_h^1 \times \tilde{H}_h^{1/2} : v(x_i) \geq 0, \forall x_i \text{ node of the partition of } \Gamma_s \text{ and } (S1, u_h|_{\Gamma} + v_h - u_0) = 0 \text{ if } d = 2\} \quad (112)$$

Note that $v_h \geq 0$ once the nodal values of v_h are ≥ 0 . Therefore, we have $D_h \subset D$. With the approximation S_h , as in (36) of S , the primal FE/BE coupling method (SP_h) reads as follows: Problem (SP_h): Find $(\hat{u}_h, \hat{v}_h) \in D_h$ such that

$$A_h(\hat{u}_h, \hat{v}_h)(u_h - \hat{u}_h, v_h - \hat{v}_h) \geq \lambda_h(u_h - \hat{u}_h, v_h - \hat{v}_h) \quad (113)$$

for all $(u_h, v_h) \in D_h$, where

$$A_h(u_h, v_h)(r_h, s_h) := DG(u_h, r_h) + (S_h(u_h|_{\Gamma} + v_h), r_h|_{\Gamma} + s_h) \quad (114)$$

and

$$\lambda_h(u_h, v_h) := L(u_h, u_h|_{\Gamma} + v_h) + (S_h u_0, u_h|_{\Gamma} + v_h) \quad (115)$$

with the discrete Steklov–Poincaré operator S_h (57).

As shown in Carstensen and Gwinner (1997), the solution $(\hat{u}_h, \hat{v}_h) \in D_h$ of (SP_h) converges for $h \rightarrow 0$ toward the solution $(\hat{u}, \hat{v}) \in D$ of (SP). Maischak (2001) presents an a posteriori error estimate on the basis of hierarchical subspace decompositions, extending the approach described in Section 2.2.2 to unilateral problems, and investigates an hp-version for the coupling method.

Next, we comment on the solvers for (SP_h) in the linear case ($q = 1$), where the matrix on the left-hand side of (113) becomes

$$A_h := \begin{pmatrix} A & B & 0 \\ B^T & C + S_{\Gamma\Gamma} & S_{\Gamma\Gamma} \\ 0 & S_{\Gamma\Gamma}^T & S_{SS} \end{pmatrix} \quad (116)$$

Here, A, B, C denote the different parts of the FEM-matrix (A belonging to the interior nodes, C belonging to all boundary nodes, and B belonging to the coupling of interior and boundary nodes), and S with its different subscripts is the Steklov–Poincaré operator acting either on whole Γ , or Γ_s , or both.

In Maischak (2001), 2-block preconditioners of the form

$$B := \begin{pmatrix} B_{ABC} & 0 \\ 0 & B_S \end{pmatrix}$$

are applied, where B_{ABC} is the symmetric V-cycle MG preconditioner $B_{MG,ABC}$ or the BPX preconditioner $B_{BPX,ABC}$

belonging to $\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$ + mass matrix, and B_S is the MG V-cycle preconditioner $B_{MG,S}$ or the BPX preconditioner $B_{BPX,S}$ belonging to S_{SS} respectively. Then, the preconditioned systems have bounded or mildly growing condition numbers, that is,

$$\kappa(B_{MG}A_h) \leq C, \quad \kappa(B_{BPX}A_h) \leq C \left(1 + \left(\log \frac{1}{h}\right)^2\right)$$

with some constant C .

From Maischak (2001), we present numerical experiments for a two-dimensional interface problem with Signorini conditions on both large sides of the L-shaped domain with vertices $(0, 0)$, $(0, 1/4)$, $(-1/4, 1/4)$, $(-1/4, -1/4)$, $(1/4, -1/4)$, and $(1/4, 0)$.

We set $q = 1$, $f = 0$, and $u_0 = r^{2/3} \sin(2/3)(\varphi - (\pi/2))$, $t_0 = (\partial/\partial n)u_0$, and $\Gamma_s = (-1/4, -1/4) \cup (-1/4, 1/4) \cup (1/4, -1/4) \cup (1/4, 0)$.

All computations are done using rectangular mesh elements with linear test and trial functions for the FEM part. We have tested the preconditioned Polyak algorithm. Note that the Polyak algorithm is a modification of the CG algorithm (see O’Leary, 1980). Preconditioners have been the MG algorithm (V-cycle, one pre- and post-smoothing step using dampened Jacobi with damping-factor 0.5) and the BPX algorithm. Tables 4, 5, and 6 give the extreme eigenvalues λ_{\min} , λ_{\max} , and the condition numbers κ for the original system, the system with multigrid preconditioner, and the system with BPX preconditioner. We note the linear growth of the condition number of the original system, the logarithmic growth of the system with BPX preconditioner, and that the condition numbers for the system with MG preconditioner are bounded (see Chapter 6, Volume 2).

5.2 Dual-mixed method

Now, we consider again problem (103) to (107), but restrict ourselves to the linear case $q = 1$.

In this section, we give a dual-mixed variational formulation of this linear Signorini contact problem in terms of a

Table 4. Extreme eigenvalues and condition numbers κ of A_h .

| N | λ_{\min} | λ_{\max} | κ |
|-----------|------------------|------------------|-----------|
| 20 + 8 | 0.1191171 | 6.8666176 | 57.645964 |
| 64 + 16 | 0.0440775 | 7.3506673 | 171.30417 |
| 232 + 24 | 0.0140609 | 7.8599693 | 558.99601 |
| 864 + 32 | 0.0040314 | 7.9613671 | 1974.8471 |
| 3288 + 40 | 0.0010835 | 7.9899020 | 7373.9834 |

Table 5. Extreme eigenvalues and condition numbers κ of B_{MG} A_h with multigrid.

| N | λ_{\min} | λ_{\max} | κ |
|-----------|------------------|------------------|-----------|
| 20 + 8 | 0.2638324 | 51.945662 | 196.88885 |
| 64 + 16 | 0.2544984 | 52.189897 | 205.06963 |
| 232 + 24 | 0.2484367 | 52.274680 | 210.41444 |
| 864 + 32 | 0.2442467 | 52.303797 | 214.14336 |
| 3288 + 40 | 0.2412018 | 52.313155 | 216.88543 |

Table 6. Extreme eigenvalues and condition numbers κ of B_{BPX} A_h with BPX.

| N | λ_{\min} | λ_{\max} | κ |
|-----------|------------------|------------------|-----------|
| 20 + 8 | 0.3812592 | 10.995214 | 28.839212 |
| 64 + 16 | 0.4267841 | 15.119538 | 35.426666 |
| 232 + 24 | 0.4458868 | 17.891084 | 40.124721 |
| 864 + 32 | 0.4544801 | 19.893254 | 43.771457 |
| 3288 + 40 | 0.4585964 | 21.465963 | 46.807965 |

convex minimization problem and an associated variational inequality.

In Maischak (2001) and Gatica, Maischak and Stephan (2003), a coupling method is proposed and analyzed for dual-mixed FE and BE for (103) to (107) using the inverse Steklov–Poincaré operator R given by

$$R := S^{-1} = V + (I + K)W^{-1}(I + K') : H^{-1/2}(\Gamma) \rightarrow H^{1/2}(\Gamma) \quad (117)$$

Note that the operator $W : H^{1/2}(\Gamma)/\mathbb{R} \rightarrow H^{-1/2}(\Gamma)$ is positive definite and therefore continuously invertible.

In order to fix the constant in $H^{1/2}(\Gamma)/\mathbb{R}$, one usually chooses the subspace of functions with integral mean zero. However, since this is not optimal from the implementational point of view, we add a least squares term to the hypersingular integral operator W . In other words, we define the functional $P : H^{1/2}(\Gamma) \rightarrow \mathbb{R}$, where $P(\phi) = \int_{\Gamma} \phi \, ds$ for all $\phi \in H^{1/2}(\Gamma)$ with adjoint P' , and set the positive definite operator

$$\tilde{W} := W + P'P : H^{1/2}(\Gamma) \rightarrow H^{-1/2}(\Gamma) \quad (118)$$

In this way, we can evaluate $u := R(t)$ for $t \in H^{-1/2}(\Gamma)$ by computing $u = (1/2)(Vt + (I + K)\phi)$, where ϕ is the solution of

$$\tilde{W}\phi = (I + K')t \quad (119)$$

Representing the solution ϕ of (119) as $\phi = \phi_0 + c_\phi$, such that $P\phi_0 = 0$ and $c_\phi \in \mathbb{R}$, and using that $(\tilde{W}\phi, 1) = ((I + K')t, 1)$, $W1 = 0$, and $K1 = -1$, we deduce that

$(P1, P\phi) = 0$, and, consequently, $c_\phi = 0$ and $\phi = \phi_0 \in H^{1/2}(\Gamma)/\mathbb{R}$ is the unique solution of $W\phi = (I + K')t$.

Therefore, we can replace W and $H^{1/2}(\Gamma)/\mathbb{R}$ by \tilde{W} and $H^{1/2}(\Gamma)$ for the discretization without mentioning it explicitly.

Next, we introduce the dual formulation $(\tilde{S}P)$ using the inverse Steklov–Poincaré operator R . To this end, we define $\tilde{\Psi} : H(\text{div}; \Omega) \rightarrow \mathbb{R} \cup \{\infty\}$ by

$$\tilde{\Psi}(q) := \frac{1}{2} \|q\|_{L^2(\Omega)^d}^2 + \frac{1}{2} (q \cdot n, R(q \cdot n)) - \frac{1}{2} (q \cdot n, R(t_0) + 2u_0) \quad (120)$$

and the subset of admissible functions by

$$\tilde{D} := \{q \in H(\text{div}; \Omega) : q \cdot n \leq 0 \text{ on } \Gamma_s, -\text{div } q = f \text{ in } \Omega\}$$

Then, the uniquely solvable problem $(\tilde{S}P)$ consists in finding $q^D \in \tilde{D}$ such that

$$\tilde{\Psi}(q^D) = \min_{q \in \tilde{D}} \tilde{\Psi}(q) \quad (121)$$

As shown in Maischak (2001) and Gatica, Maischak and Stephan (2003), problem $(\tilde{S}P)$ is equivalent to the original Signorini contact problem (103) to (107) with $q = 1$.

Next, we want to introduce a saddle-point formulation of $(\tilde{S}P)$ and define $\mathcal{H} : H(\text{div}; \Omega) \times L^2(\Omega) \times H_0^{1/2}(\Gamma_s) \rightarrow \mathbb{R} \cup \{\infty\}$ as

$$\mathcal{H}(p, v, \mu) := \tilde{\Psi}(p) + \int_{\Omega} v \, \text{div } p \, dx + \int_{\Omega} f v \, dx + (p \cdot n, \mu)_{\Gamma_s} \quad (122)$$

for all $(p, v, \mu) \in H(\text{div}; \Omega) \times L^2(\Omega) \times H_0^{1/2}(\Gamma_s)$, and consider the subset of admissible functions

$$\tilde{H}_0^{1/2}(\Gamma_s) := \{\mu \in H_0^{1/2}(\Gamma_s) : \mu \geq 0\} \quad (123)$$

Then the desired saddle-point problem (M) reads as follows:

Problem (M): Find $(\hat{q}, \hat{u}, \hat{\lambda}) \in H(\text{div}; \Omega) \times L^2(\Omega) \times \tilde{H}_0^{1/2}(\Gamma_s)$ such that

$$\mathcal{H}(\hat{q}, u, \lambda) \leq \mathcal{H}(\hat{q}, \hat{u}, \hat{\lambda}) \leq \mathcal{H}(q, \hat{u}, \hat{\lambda}) \quad \forall (q, u, \lambda) \in H(\text{div}; \Omega) \times L^2(\Omega) \times \tilde{H}_0^{1/2}(\Gamma_s) \quad (124)$$

which is equivalent (see Ekeland and Temam, 1974) to finding a solution $(\hat{q}, \hat{u}, \hat{\lambda}) \in H(\text{div}; \Omega) \times L^2(\Omega) \times$

$\tilde{H}_{00}^{1/2}(\Gamma_s)$ of the following variational inequality:

$$a(\hat{q}, q) + b(q, \hat{u}) + d(q, \hat{\lambda}) = (q \cdot n, r) \quad \forall q \in H(\operatorname{div}; \Omega) \quad (125)$$

$$b(\hat{q}, u) = - \int_{\Omega} f u \, dx \quad \forall u \in L^2(\Omega) \quad (126)$$

$$d(\hat{q}, \lambda - \hat{\lambda}) \leq 0 \quad \forall \lambda \in \tilde{H}_{00}^{1/2}(\Gamma_s) \quad (127)$$

where

$$a(p, q) = 2 \int_{\Omega} p \cdot q \, dx + (q \cdot n, R(p \cdot n)) \quad \forall p, q \in H(\operatorname{div}; \Omega) \quad (128)$$

$$b(q, u) = \int_{\Omega} u \operatorname{div} q \, dx \quad \forall q, u \in H(\operatorname{div}; \Omega) \times L^2(\Omega) \quad (129)$$

$$d(q, \lambda) = (q \cdot n, \lambda)_{\Gamma_s} \quad \forall (q, \lambda) \in H(\operatorname{div}; \Omega) \times H_{00}^{1/2}(\Gamma_s) \quad (130)$$

and $r = R(t_0) + 2u_0$.

Now, we define the bilinear form

$$B(q, (u, \lambda)) = b(q, u) + d(q, \lambda) \quad \forall (q, u, \lambda) \in H(\operatorname{div}; \Omega) \times L^2(\Omega) \times H_{00}^{1/2}(\Gamma_s) \quad (131)$$

and observe that the above variational inequality can be written as

$$\begin{aligned} a(\hat{q}, q) + B(q, (\hat{u}, \hat{\lambda})) &= (q \cdot n, r) \quad \forall q \in H(\operatorname{div}; \Omega) \\ B(\hat{q}, (u - \hat{u}, \lambda - \hat{\lambda})) &\leq - \int_{\Omega} f(u - \hat{u}) \, dx \quad \forall (u, \lambda) \in L^2(\Omega) \times H_{00}^{1/2}(\Gamma_s) \end{aligned} \quad (132)$$

The dual problem $(\tilde{S}\tilde{P})$ and the saddle-point problem (M) are related to each other as follows.

Theorem 14. *The dual problem $(\tilde{S}\tilde{P})$ is equivalent to the mixed dual variational inequality (M) . More precisely, (i) If $(\hat{q}, \hat{u}, \hat{\lambda}) \in H(\operatorname{div}; \Omega) \times L^2(\Omega) \times H_{00}^{1/2}(\Gamma_s)$ is a saddle point of \mathcal{H} in $H(\operatorname{div}; \Omega) \times L^2(\Omega) \times H_{00}^{1/2}(\Gamma_s)$, then $\hat{q} = \nabla \hat{u}$, $\hat{u} = 1/2 R(\hat{q} \cdot n) + u_0$ on Γ_s , $\hat{\lambda} = -(1/2) R(\hat{q} \cdot n - t_0) + u_0 - \hat{u}$ on Γ_s , and $\hat{q} \in \tilde{D}$ is the solution of problem $(\tilde{S}\tilde{P})$.*

(ii) Let $q^D \in \tilde{D}$ be the solution of $(\tilde{S}\tilde{P})$, and define $\hat{\lambda} := -(1/2) R(q^D \cdot n - t_0) + u_0 - \hat{u}$ on Γ_s , where $\hat{u} \in H^1(\Omega)$ is the unique solution of the Neumann problem: $-\Delta \hat{u} = f$ in Ω , $\partial \hat{u} / \partial n = q^D \cdot n$ on Γ , such that $(\mu, \hat{u} + 1/2 R(q^D \cdot n - t_0) - u_0) \geq 0$ for all $\mu \in H^{-1/2}(\Gamma)$ with $\mu \leq -q^D \cdot n$ on

Γ_s . Then, $(q^D, \hat{u}, \hat{\lambda})$ is a saddle point of \mathcal{H} in $H(\operatorname{div}; \Omega) \times L^2(\Omega) \times H_{00}^{1/2}(\Gamma_s)$.

Next, we deal with the numerical approximation for problem $(\tilde{S}\tilde{P})$ by using mixed FE in Ω and BE on Γ_s , as given in Maischak (2001) and Gatica, Maischak and Stephan (2003). For simplicity, we assume that Γ_s and Γ_n are polygonal (i.e. piecewise straight lines) for $d = 2$ or piecewise hyperplanes for $d \geq 3$.

Let $(T_h)_{h \in I}$ be a family of regular triangulations of the domain Ω by triangles/tetrahedrons T of diameter h_T such that $h := \max\{h_T : T \in T_h\}$. We denote by ρ_T the diameter of the inscribed circle/sphere in T , and assume that there exists a constant $\kappa > 0$ such that for any h and for any T in T_h , the inequality $(h_T/\rho_T) \leq \kappa$ holds. Moreover, we assume that there exists a constant $C > 0$ such that for any h and for any triangle/tetrahedron T in T_h with $T \cap \partial\Omega$ is a whole edge/face of T , there holds $|T \cap \partial\Omega| \geq C h^{d-1}$, where $|T \cap \partial\Omega|$ denotes the length/area of $T \cap \partial\Omega$. This means that the family of triangulations is uniformly regular near the boundary.

We also assume that all the points/curves in $\tilde{\Gamma}_s \cap \tilde{\Gamma}_n$ become vertices/edges of T_h for all $h > 0$. Then, we denote by \mathcal{E}_h the set of all edges/faces e of T_h and put $\mathcal{E}_h := \{e \in \mathcal{E}_h : e \subset \Gamma\}$. Further, let $(\tau_k)_{k \in I}$ be a family of independent regular triangulations of the boundary part Γ_s by line segments/triangles Δ of diameter h_Δ such that $h := \max\{h_\Delta : \Delta \in \tau_k\}$.

We take $I \subset (0, \infty)$ with $0 \in I$, and choose a family of finite-dimensional subspaces $(X_{h,h})_{h \in I, h \in I} = (L_h \times H_h \times H_h^{-1/2} \times H_h^{1/2} \times H_h^{1/2})$ of $X = L^2(\Omega) \times H(\operatorname{div}; \Omega) \times H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)/R \times H_{00}^{1/2}(\Gamma_s)$, subordinated to the corresponding triangulations, with $H_h^{-1/2}$ being the restriction of H_h on Γ_s , and we assume that the following approximation property holds

$$\lim_{h \rightarrow 0} \left\{ \inf_{(u_h, q_h, \psi_h, \phi_h, \lambda_h) \in X_{h,h}} \| (u, q, \psi, \phi, \lambda) - (u_h, q_h, \psi_h, \phi_h, \lambda_h) \|_X \right\} = 0 \quad (134)$$

for all $(u, q, \psi, \phi, \lambda) \in X$. In addition, we assume that the divergence of the functions in H_h belong to L_h , that is,

$$\{\operatorname{div} q_h : q_h \in H_h\} \subseteq L_h \quad (135)$$

Also, the subspaces $(L_h, H_h^{1/2})$ and H_h are supposed to verify the usual discrete Babuška–Brezzi condition, which

means that there exists $\beta^* > 0$ such that

$$\inf_{\substack{(u_h, \lambda_h) \in H_h^{1/2} \times H_h^{1/2} \\ (u_h, \lambda_h) \neq 0}} \sup_{\substack{q_h \in H(\operatorname{div}; \Omega) \\ (q_h, \lambda_h) \neq 0}} \frac{B(q_h, (u_h, \lambda_h))}{|q_h|_{H(\operatorname{div}; \Omega)} \| (u_h, \lambda_h) \|_{L^2(\Omega) \times H_{00}^{1/2}(\Gamma_s)}} \geq \beta^* \quad (136)$$

Now, for $h, \tilde{h} \in I$, let $j_h : H_h \hookrightarrow H(\operatorname{div}; \Omega)$, $k_h : H_h^{-1/2} \hookrightarrow H^{-1/2}(\Gamma)$ and $l_h : H_h^{1/2} \hookrightarrow H^{1/2}(\Gamma)/R$ denote the canonical imbeddings with their corresponding duals j_h^* , k_h^* and l_h^* .

In order to approximate R , we define the discrete operators

$$R_h := j_h^* \gamma^* R \gamma j_h \quad \text{and} \quad \tilde{R}_h := j_h^* \gamma^* V \gamma j_h + j_h^* \gamma^* \times (I + K) l_h (l_h^* W l_h)^{-1} l_h^* (I + K') \gamma j_h$$

where $\gamma : H(\operatorname{div}; \Omega) \rightarrow H^{-1/2}(\Gamma)$ is the trace operator yielding the normal component of functions in $H(\operatorname{div}; \Omega)$.

We remark that the computation of \tilde{R}_h requires the numerical solution of a linear system with a symmetric positive definite matrix $W_h := l_h^* W l_h$. In general, there holds $\tilde{R}_h \neq R_h$ because \tilde{R}_h is a Schur complement of discretized matrices, while R_h is a discretized Schur complement of operators.

Then, in order to approximate the solution of problem (M) , we consider the following nonconforming Galerkin scheme (M_h) :

Problem (M_h) : Find $(\hat{q}_h, \hat{u}_h, \hat{\lambda}_h) \in H_h \times L_h \times H_{s,h}^{1/2}$ such that

$$a_h(\hat{q}_h, q_h) + b(q_h, \hat{u}_h) + d(q_h, \hat{\lambda}_h) = (q_h \cdot n, r_h) \quad \forall q_h \in H_h \quad (137)$$

$$b(\hat{q}_h, u_h) = - \int_{\Omega} f u_h \, dx \quad \forall u_h \in L_h \quad (138)$$

$$d(\hat{q}_h, \lambda_h - \hat{\lambda}_h) \leq 0 \quad \forall \lambda_h \in H_{s,h}^{1/2} \quad (139)$$

where

$$H_{s,h}^{1/2} := \{\mu \in H_s^{1/2} : \mu \geq 0\} \quad (140)$$

$$a_h(p, q) = 2 \int_{\Omega} p \cdot q \, dx + (q \cdot n, \tilde{R}_h(p \cdot n)) \quad \forall p, q \in H_h \quad (141)$$

$$b(q, u) = \int_{\Omega} u \operatorname{div} q \, dx \quad \forall (q, u) \in H_h \times L_h \quad (142)$$

$$d(q, \lambda) = (q \cdot n, \lambda)_{\Gamma_s} \quad \forall (q, \lambda) \in H_h \times H_{s,h}^{1/2} \quad (143)$$

and

$$r_h := k_h^* [(V + (I + K) l_h (l_h^* W l_h)^{-1} l_h^* (I + K')) t_0 + 2u_0]$$

Note that the nonconformity of problem (M_h) arises from the bilinear form $a_h(\cdot, \cdot)$ approximating $a(\cdot, \cdot)$.

There holds the following a priori error estimate (see Maischak (2001) and Gatica, Maischak and Stephan (2003)) yielding convergence for the solution of the nonconforming Galerkin scheme (M_h) to the weak solution of (M) and therefore to the weak solution of the original Signorini contact problem owing to the equivalence result of Theorem 14.

Theorem 15. *Let $(\hat{q}, \hat{u}, \hat{\lambda})$ and $(\hat{q}_h, \hat{u}_h, \hat{\lambda}_h)$ be the solutions of problems (M) and (M_h) respectively. Define $\hat{\phi} := W^{-1}(I + K')(\hat{q} \cdot n)$ and $\hat{\phi}_h := W^{-1}(I + K')\hat{q}_h$. Then, there exists $c > 0$, independent of h and \tilde{h} , such that the following Cea type estimate holds:*

$$\begin{aligned} \|\hat{q} - \hat{q}_h\|_{H(\operatorname{div}; \Omega)} + \|\hat{u} - \hat{u}_h\|_{L^2(\Omega)} + \|\hat{\lambda} - \hat{\lambda}_h\|_{H_{00}^{1/2}(\Gamma_s)} \\ \leq c \left\{ \inf_{q_h \in H_h} \|\hat{q} - q_h\|_{H(\operatorname{div}; \Omega)} + \inf_{u_h \in L_h} \|\hat{u} - u_h\|_{L^2(\Omega)} \right. \\ \left. + \inf_{\lambda_h \in H_{s,h}^{1/2}} \|\hat{\lambda} - \lambda_h\|_{H_{00}^{1/2}(\Gamma_s)} + \inf_{\phi_h \in H_h^{1/2}} \|\hat{\phi} - \phi_h\|_{H^{1/2}(\Gamma)/R} \right. \\ \left. + \inf_{\phi_h \in H_h^{1/2}} \|\hat{\phi}_0 - \phi_h\|_{H^{1/2}(\Gamma)/R} \right\} \quad (144) \end{aligned}$$

The proof of Theorem 15 uses besides (134) and (135) the discrete Babuška–Brezzi condition (136). A suitable choice for FE and BE spaces are L_h , the set of piecewise constant functions, H_h , the space of $H(\operatorname{div}; \Omega)$ conforming RT elements of order zero, and $H_{s,h}^{1/2}$, the set of continuous piecewise linear, nonnegative functions of the partition τ_k of Γ_s .

Next, we present an a posteriori error estimate with residual type estimator, which is given in Maischak (2001).

Theorem 16. *Let $d = 2$. There exists $C > 0$, independent of h, \tilde{h} , such that*

$$\begin{aligned} \|\hat{q} - \hat{q}_h\|_{H(\operatorname{div}; \Omega)} + \|\hat{u} - \hat{u}_h\|_{L^2(\Omega)} + \|\hat{\lambda} - \hat{\lambda}_h\|_{H_{00}^{1/2}(\Gamma_s)} \\ \leq C \left(\sum_{T \in \mathcal{T}_h} \eta_T^2 \right)^{1/2} \quad (145) \end{aligned}$$

where, for any triangle $T \in \mathcal{T}_h$, we define

$$\begin{aligned} \eta_T^2 = & \|f + \operatorname{div} \hat{q}_h\|_{L^2(T)}^2 \\ & + h_T^2 \|\operatorname{curl}(\hat{q}_h)\|_{L^2(T)}^2 + h_T^2 \|\hat{q}_h\|_{L^2(T)}^2 \\ & + \sum_{e \in \mathcal{E}(T) \cap \Omega_h(\Gamma)} h_e \|\xi_h - u_0 + \hat{\lambda}_h - s_h\|_{L^2(e)}^2 \\ & + \sum_{e \in \mathcal{E}(T) \cap \Omega_h(\Omega)} h_e \|\hat{q}_h \cdot \vec{n}\|_{L^2(e)}^2 \\ & + \sum_{e \in \mathcal{E}(T) \cap \Omega_h(\Gamma)} h_e \left\| \hat{q}_h \cdot \vec{t} + \frac{d}{ds} (\xi_h - u_0) \right\|_{L^2(e)}^2 \\ & + \sum_{e \in \mathcal{E}(T) \cap \Omega_h(\Gamma)} h_e \|\xi_h\|_{L^2(e)}^2 + \sum_{e \in \mathcal{E}(T) \cap \Omega_h(\Omega)} h_e \left\| \frac{d}{ds} \lambda_h \right\|_{L^2(e)}^2 \end{aligned}$$

with s_h being the L^2 -projection of $\xi_h - u_0 + \hat{\lambda}_h$ onto the space of piecewise constant functions on $T_h \cap \Gamma$, and $\mathcal{E}(T)$ being the edges of T , using

$$\begin{aligned} \xi_h &:= V(\hat{q}_h \cdot n - t_0) - (I + K)\hat{\phi}_h \\ \zeta_h &:= W\hat{\phi}_h + (I + K^*)(\hat{q}_h \cdot n - t_0) \\ \hat{\phi}_h &:= (I_h^* W I_h)^{-1} I_h^* (I + K^*) \gamma_{j_h}(\hat{q}_h \cdot n) \end{aligned}$$

In order to solve the discretized saddle-point problem (M_h) , Maischak (2001) and Gatica, Maischak and Stephan (2003) propose a modified Uzawa algorithm that utilizes the equation for the Lagrange multiplier λ . For this purpose, we first introduce the operators $P_h: H_{s,h}^{1/2} \rightarrow H_{s,h}^{1/2}$ and $\Phi: H(\operatorname{div}; \Omega) \rightarrow H_{s,h}^{1/2} \subset H_{00}^{1/2}(\Gamma_s)$ that are defined as

$$(P_h \lambda - \lambda, \lambda_h - P_h \lambda)_{H_{00}^{1/2}(\Gamma_s)} \geq 0 \quad \forall \lambda_h \in H_{s,h}^{1/2}$$

and

$$d(q, \lambda_h) = (\lambda_h, \Phi(q))_{H_{00}^{1/2}(\Gamma_s)} \quad \forall \lambda_h \in H_{s,h}^{1/2} \quad (146)$$

with given $\lambda \in H_{s,h}^{1/2}$ and $q \in H(\operatorname{div}; \Omega)$.

Now, the modified Uzawa algorithm is formulated as follows.

1. Choose an initial $\lambda^{(0)} \in H_{s,h}^{1/2}$.
2. Given $\lambda^{(n)} \in H_{s,h}^{1/2}$, find $q^{(n)}, u^{(n)} \in H_h \times L_h$ such that

$$a_h(q^{(n)}, q_h) + b(q_h, u^{(n)}) = (f_h, q_h \cdot n) - d(q_h, \lambda^{(n)}) \quad \forall q_h \in H_h \quad (147)$$

$$b(q^{(n)}, u_h) = -(f, u_h)_{L^2(\Omega)} \quad \forall u_h \in L_h \quad (148)$$

3. Compute $\lambda^{(n+1)}$ by

$$\lambda^{(n+1)} = P_h(\lambda^{(n)} + \delta \Phi(q^{(n)})) \quad (149)$$

4. Check for some stopping criterion and if it is not fulfilled, then go to step 2.

The convergence of this algorithm is established in the following theorem (Maischak, 2001 and Gatica, Maischak and Stephan, 2003).

Theorem 17. Let $\delta \in]0, 2[$ and consider any initial value $\lambda_0 \in H_{s,h}^{1/2}$. Then, the modified Uzawa algorithm converges toward the solution of the discrete problem (M_h) .

We remark that the operators P_h and Φ are defined with respect to the scalar product of the Sobolev space $H_{00}^{1/2}(\Gamma_s)$, which is not practical from the computational point of view. Fortunately, in the convergence proof we only need that the norm induced by the scalar product is equivalent to the $H_{00}^{1/2}(\Gamma_s)$ -norm. Therefore, we can use the bilinear form $(W \cdot, \cdot)$ instead of the scalar product $(\cdot, \cdot)_{H_{00}^{1/2}(\Gamma_s)}$. Then, the computation of the projection $P_h \lambda$ now leads to the following variational inequality: find $P_h \lambda \in H_{s,h}^{1/2}$ such that

$$\begin{aligned} (WP_h \lambda, \lambda_h - P_h \lambda) &\geq (W \lambda, \lambda_h - P_h \lambda) \\ \forall \lambda_h &\in H_{s,h}^{1/2} \quad (150) \end{aligned}$$

Similarly, the computation of the operator Φ is done by solving a linear equation: find $\Phi(q) \in H_{s,h}^{1/2}$ such that

$$(W\Phi(q), \lambda_h) = d(q, \lambda_h) \quad \forall \lambda_h \in H_{s,h}^{1/2} \quad (151)$$

Both systems are small compared to the total size of the problem because they are only defined on the Signorini part Γ_s of the interface Γ . Applying (151) to (150), we obtain for $\lambda = \lambda^{(n)} + \delta \Phi(q^{(n)})$,

$$\begin{aligned} (W \lambda, \lambda_h - P_h \lambda) &= (W \lambda^{(n)}, \lambda_h - P_h \lambda) \\ &\quad + \delta (W \Phi(q^{(n)}), \lambda_h - P_h \lambda) \\ &= (W \lambda^{(n)}, \lambda_h - P_h \lambda) \\ &\quad + \delta d(q^{(n)}, \lambda_h - P_h \lambda) \end{aligned}$$

and hence the explicit solution of (151) is avoided.

Finally, we present iteration numbers for the Uzawa algorithm applied to the above dual-mixed FE/BE coupling method for the example in Section 5.1 with the above choice of FE/BE spaces given after Theorem 15.

Table 7 gives the numbers of outer iterations for the Uzawa algorithm with $\delta = 1.3$. We notice that the numbers

Table 7. Iteration numbers for the Uzawa algorithm.

| dim L_h | dim H_h | dim $H_{s,h}^{1/2}$ | Iteration |
|-----------|-----------|---------------------|-----------|
| 12 | 32 | 3 | 47 |
| 48 | 112 | 7 | 48 |
| 192 | 416 | 15 | 47 |
| 768 | 1600 | 31 | 47 |
| 3072 | 6272 | 63 | 48 |
| 12288 | 24832 | 127 | 47 |
| 49152 | 98816 | 255 | 46 |

of outer iterations are nearly independent of the problem size. The inner linear systems are solved with the GMRES algorithm (see Chapter 9, this Volume).

6 APPLICATIONS

6.1 Symmetric coupling of standard finite elements and boundary elements for Hencky-elasticity

In the following, we consider an interface problem from nonlinear elasticity in which in a three-dimensional bounded domain Ω_1 , with a hole, the material is nonlinear elastic obeying the Hencky-von Mises stress-strain relation, and it is linear elastic in a surrounding unbounded exterior region Ω_2 . The boundaries Γ_u and Γ of Ω_1 are assumed to be Lipschitz continuous, where Γ_u denotes the boundary of the hole and Γ is the interface boundary between Ω_1 and Ω_2 . We assume the nonlinear Hencky-von Mises stress-strain relation of the form

$$\sigma = (k - \frac{2}{3}\mu(\gamma))\mathbf{I} \cdot \operatorname{div} \mathbf{u}_1 + 2\mu(\gamma)\epsilon$$

where σ and $\epsilon = 1/2(\nabla \mathbf{u}^T + \nabla \mathbf{u})$ denote the (Cauchy) stresses and the (linear Green) strain respectively (see Nečas and Hlaváček, 1981; Nečas, 1986; Zeidler, 1988). Then, if we define

$$\begin{aligned} P_1(\mathbf{u}_1)_i &:= \frac{\partial}{\partial x_i} \left(k - \frac{2}{3}\mu(\gamma(\mathbf{u}_1)) \right) \\ &\quad \cdot \operatorname{div} \mathbf{u}_1 + \sum_{j=1}^3 \frac{\partial}{\partial x_j} \mu(\gamma(\mathbf{u}_1)) \epsilon_{ij}(\mathbf{u}_1) \end{aligned}$$

for $i = 1, 2, 3$, the equilibrium condition $\operatorname{div} \sigma + F = 0$ gives

$$P_1(\mathbf{u}_1) = F \quad \text{in } \Omega_1 \quad (152)$$

Here, the bulk modulus k and the function $\mu(\gamma)$ in P_1 satisfy (cf. e.g. Nečas, 1986)

$$0 < \bar{\mu}_0 \leq \mu(\gamma) \leq \frac{3}{2}k$$

$$0 < \bar{\mu}_1 \leq \mu + 2\gamma \frac{d\mu}{d\gamma} \leq \bar{\mu}_2 < \infty$$

where $\bar{\mu}_0, \bar{\mu}_1, \bar{\mu}_2$ are constants, and

$$\begin{aligned} \gamma(\mathbf{u}_1) &= \sum_{i,j=1}^3 \left(\epsilon_{ij} - \delta_{ij} \frac{1}{3} \cdot \operatorname{div} \mathbf{u}_1 \right)^2 \\ \epsilon_{ij} &= \frac{1}{2} \left(\frac{\partial \mathbf{u}_{1j}}{\partial x_i} + \frac{\partial \mathbf{u}_{1i}}{\partial x_j} \right) \end{aligned}$$

In a surrounding unbounded exterior region Ω_2 , we consider the homogeneous Lamé system describing linear isotropic elastic material, with the Lamé constants $\mu_2 > 0$, $3\lambda_2 + 2\mu_2 > 0$,

$$\begin{aligned} P_2(\mathbf{u}_2) &= -\mu_2 \Delta \mathbf{u}_2 \\ &\quad - (\lambda_2 + \mu_2) \operatorname{grad} \operatorname{div} \mathbf{u}_2 = 0 \quad \text{in } \Omega_2 \quad (153) \end{aligned}$$

In Costabel and Stephan (1990), we consider the following interface problem (see also Gatica and Hsiao, 1995): For a given vector field F in Ω_1 , find the vector fields \mathbf{u}_j in Ω_j ($j = 1, 2$) satisfying $\mathbf{u}_1|_{\Gamma_u} = 0$, the differential equations (152) and (153), the interface conditions

$$\mathbf{u}_1 = \mathbf{u}_2, \quad T_1(\mathbf{u}_1) = T_2(\mathbf{u}_2) \quad \text{on } \Gamma \quad (154)$$

and the regularity condition at infinity

$$\mathbf{u}_2 = \mathcal{O}\left(\frac{1}{|x|}\right) \quad \text{as } |x| \rightarrow \infty \quad (155)$$

Here, with $\mu_1 = \mu(\gamma(\mathbf{u}_1))$, $\lambda_1 = k - (2/3)\mu(\gamma(\mathbf{u}_1))$, the tractions are given by

$$T_j(\mathbf{u}_j) = 2\mu_j \partial_n \mathbf{u}_j + \lambda_j n \operatorname{div} \mathbf{u}_j + \mu_j n \times \operatorname{curl} \mathbf{u}_j \quad (156)$$

and $\partial_n \mathbf{u}_j$ is the derivative with respect to the outer normal on Γ .

We are interested in solutions \mathbf{u}_j of (152) to (155), which belong to $(H_{\infty}^1(\Omega_j))^3$, that is, which are of finite energy. A variational formulation is obtained as in Costabel and Stephan (1990). An application of the first Green formula to (152) yields

$$\int_{\Omega_1} P_1(\mathbf{u}_1) w \, dx = \Phi_1(\mathbf{u}_1, w) - \int_{\Gamma} T_1(\mathbf{u}_1) w \, ds \quad (157)$$

for all $w \in [H^1(\Omega_1)]^3$, where

$$\Phi_1(u_1, w) := \int_{\Omega_1} \left\{ \left(k - \frac{2}{3} \mu(\gamma(u_1)) \right) \operatorname{div} u_1 \operatorname{div} w + \sum_{i,j=1}^3 2\mu(\gamma(u_1)) \epsilon_{ij}(u_1) \epsilon_{ij}(w) \right\} dx \quad (158)$$

On the other hand, the solution u_2 of (153) is given by the Somigliana representation formula for $x \in \Omega_2$:

$$u_2(x) = \int_{\Gamma} (T_2(x, y) v_2(y) - G_2(x, y) \phi_2(y)) ds(y) \quad (159)$$

where $v_2 = u_2$, $\phi_2 = T_2(u_2)$ on Γ , and the fundamental solution $G_2(x, y)$ of $P_2 u_2 = 0$ is the 3×3 matrix function

$$G_2(x, y) = \frac{\lambda_2 + 3\mu_2}{8\pi\mu_2(\lambda_2 + 2\mu_2)} \times \left\{ \frac{1}{|x - y|} I + \frac{\lambda_2 + \mu_2}{\lambda_2 + 3\mu_2} \frac{(x - y)(x - y)^T}{|x - y|^3} \right\}$$

with the unit matrix I and $T_2(x, y) = T_{2,y}(G_2(x, y))^T$, where superscript T denotes transposition. Taking Cauchy data in (159), that is, boundary values and tractions on Γ for $x \rightarrow \Gamma$, we obtain a system of boundary integral equations on Γ ,

$$v_2 = (I + K)v_2 - V\phi_2 \text{ and } 2\phi_2 = -Wv_2 + (I - K')\phi_2 \quad (160)$$

with the single-layer potential V , a weakly singular boundary integral operator, the double-layer potential K and its dual K' , strongly singular operators, and the hypersingular operator W defined as

$$\begin{aligned} V\phi_2(x) &= 2 \int_{\Gamma} G_2(x, y) \phi_2(y) ds(y) \\ K v_2(x) &= 2 \int_{\Gamma} T_2(x, y) v_2(y) ds(y) \\ K' \phi_2(x) &= 2 T_{2,x} \int_{\Gamma} G_2(x, y) \phi_2(y) ds(y) \\ W v_2(x) &= -2 T_{2,x} \int_{\Gamma} T_2(x, y) v_2(y) ds(y) \end{aligned}$$

As in interface problems for purely linear equations (cf. Section 2), we obtain a variational formulation for the interface problem (152) to (155) by inserting a weak form of the boundary integral equations (160) on Γ into the weak form (157) and making use of the interface conditions (154), that is, $T_2 = T_1$, $\phi_2 = \phi_1$ and $v_2 = u_1 = u$.

This yields the following variational problem: for given $F \in L^2(\Omega_1)^3$, find $u \in H^1(\Omega_1)^3$, $\phi \in H^{-1/2}(\Gamma)^3$ such that

$u|_{\Gamma_+} = 0$ and

$$b(u, \phi; w, \psi) = 2 \int_{\Omega_1} F \cdot w dx \quad (161)$$

for all $(w, \psi) \in H^1(\Omega_1)^3 \times H^{-1/2}(\Gamma)^3$. Here, with the form $\Phi_1(\cdot, \cdot)$ in (158) and the brackets (\cdot, \cdot) denoting the extended L^2 -duality between the trace space $H^{1/2}(\Gamma)^3$ and its dual $H^{-1/2}(\Gamma)^3$, we define

$$b(u, \phi; w, \psi) := 2\Phi_1(u, w) + (w, Wu) - (w, (I - K')\phi) - ((I - K)u, \psi) - (\psi, V\phi) \quad (162)$$

Theorem 18. For $F \in L^2(\Omega_1)^3$, there exists exactly one solution $u \in H^1(\Omega_1)^3$, $\phi \in H^{-1/2}(\Gamma)^3$ of (161), yielding a solution of the interface problem (152) to (155) with $u_1 = u$ in Ω_1 and u_2 given by (159) in Ω_2 .

The proof in Costabel and Stephan (1990) is based on the fact that the C^2 -functional,

$$\begin{aligned} J_1(u, \phi) &:= 2A(u) + \frac{1}{2}(u, Wu) \\ &\quad - 2 \int_{\Omega_1} F u dx + (\phi, (K - I)u) - \frac{1}{2}(\phi, V\phi) \\ A(u) &:= \int_{\Omega_1} \left\{ \frac{1}{2} k |\operatorname{div} u|^2 + \int_0^{\gamma(u)} \mu(t) dt \right\} dx \quad (163) \end{aligned}$$

has a unique saddle point $u \in H^1(\Omega_1)^3$, $\phi \in H^{-1/2}(\Gamma)^3$. The two-dimensional case, treated in Carstensen, Funken and Stephan (1997) requires minor modifications only.

A key role is played by the following properties of the functional

$$J_0(u) = A(u) - \int_{\Omega_1} F u dx$$

of the single-layer potential operator V and of the hypersingular operator W : J_0 is strictly convex, that is, there exists $\lambda, \tilde{\lambda} > 0$ such that for all $u, w \in H(\Omega_1)^3$, the second Gateaux derivative satisfies

$$\lambda \|w\|_{1,\Omega_1}^2 \leq D^2 J_0(u)[w; w] \leq \tilde{\lambda} \|w\|_{1,\Omega_1}^2 \quad (164)$$

There exists $\gamma > 0$ such that for all $\phi \in H^{-1/2}(\Gamma)^3$,

$$(\phi, V\phi) \geq \gamma \|\phi\|_{-1/2,\Gamma}^2 \quad (165)$$

and

$$(v, Wv) \geq 0 \text{ for all } v \in H^{1/2}(\Gamma)^3 \quad (166)$$

For (164) see Nečas (1986); for (165) and (166) see Costabel and Stephan (1990). The saddle-point property of the function J_1 in (163) is typical of the symmetric coupling of FE and BE.

Given finite-dimensional subspaces $X_M \times Y_N$ of $H^1(\Omega_1)^3 \times H^{-1/2}(\Gamma)^3$, the Galerkin solution $(u_M, \phi_M) \in X_M \times Y_N$ is the unique saddle point of the functional J_1 on $X_M \times Y_N$; the Galerkin scheme for (161) reads as follows: Given $F \in L^2(\Omega_1)^3$ find $u_M \in X_M$ and $\phi_M \in Y_N$ such that for all $w \in X_M$ and $\psi \in Y_N$,

$$b(u_M, \phi_M; w, \psi) = 2 \int_{\Omega_1} F \cdot w dx \quad (167)$$

The following theorem from Costabel and Stephan (1990) states the quasioptimal convergence in the energy norm for any conforming Galerkin scheme.

Theorem 19. There exists exactly one solution $(u_M, \phi_M) \in X_M \times Y_N$ of the Galerkin equations (167). There exists a constant C independent of X_M and Y_N such that

$$\begin{aligned} \|u - u_M\|_{H^1(\Omega_1)^3} + \|\phi - \phi_M\|_{H^{-1/2}(\Gamma)^3} \\ \leq C \left\{ \inf_{w \in X_M} \|u - w\|_{H^1(\Omega_1)^3} + \inf_{\psi \in Y_N} \|\phi - \psi\|_{H^{-1/2}(\Gamma)^3} \right\} \end{aligned} \quad (168)$$

where $(u, \phi) \in H^1(\Omega_1)^3 \times H^{-1/2}(\Gamma)^3$ is the exact solution of the variational problem (161).

The Galerkin solution $(u_M, \phi_M) \in X_M \times Y_N$ of (167) is the unique saddle point of the functional J_1 on $X_M \times Y_N$, that is, $DJ_1(u_M, \phi_M)(w; \psi) = 0$ for all $(w, \psi) \in X_M \times Y_N$.

Note that the symmetric coupling procedure was applied above to a nonlinear strongly monotone operator P_1 , but we mention that it can also be applied to some other nonlinear elasticity, viscoplasticity, and plasticity problems with hardening; see Carstensen (1993, 1994, 1996a) and Carstensen and Stephan (1995c). Numerical experiments with the h-version for some model problems are described in Stephan (1992).

6.2 Coupling of dual-mixed finite elements and boundary elements for plane elasticity

Following Gatica, Heuer and Stephan (2001), we consider the coupling of dual-mixed FE and BE to solve a mixed Dirichlet-Neumann problem of plane elasticity. We derive an a posteriori error estimate that is based on the solution of local Dirichlet problems and on a residual term defined on the coupling interface.

Let $\Omega = \Omega_F \cup \Gamma \cup \Omega_B$ be a polygonal domain in \mathbb{R}^2 with boundary $\partial\Omega = \Gamma_N \cup \Gamma_D$, where Γ_D is not empty and $\Omega_B \cap (\Gamma_N \cup \Gamma_D) = \emptyset$ (see Figure 3). For a given body load

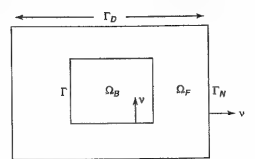


Figure 3. Geometry of the problem.

f on Ω , vanishing on Ω_B , and a given displacement g on Γ_D , we consider the linear elasticity problem

$$\operatorname{div} \sigma = -f \quad \text{in } \Omega \quad (169a)$$

$$\sigma = C e(u) \quad \text{in } \Omega \quad (169b)$$

$$u = g \quad \text{on } \Gamma_D \quad (169c)$$

$$\sigma \cdot v = 0 \quad \text{on } \Gamma_N \quad (169d)$$

Here, u is the displacement field, $e(u) := 1/2(\nabla u + (\nabla u)^T)$ is the strain tensor, and v is the outward unit normal of Ω_F . The elasticity tensor C describes the stress-strain relationship. In the simplest case, we have $\sigma = \lambda \operatorname{tr} e(u) I_2 + 2\mu e(u)$ where λ and μ are the Lamé coefficients, I_2 denotes the identity matrix in $\mathbb{R}^{2 \times 2}$, and $\operatorname{tr}(\tau) := \sum_{i=1}^2 \tau_{ii}$ for $\tau := (\tau_{ij}) \in \mathbb{R}^{2 \times 2}$. We assume that the Lamé coefficients are constant on Ω_B .

In the following, we discretize the problem (169) by coupled dual-mixed FE and BE. Below, this allows for representing u on Ω_B by pure boundary integral operators acting on Γ .

To derive an appropriate variational formulation, we define the tensor spaces

$$\begin{aligned} [L^2(\Omega_F)]^{2 \times 2} &:= \left\{ \tau = \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix}; \right. \\ &\quad \left. \tau_{ij} \in L^2(\Omega_F), i, j = 1, 2 \right\} \end{aligned}$$

and

$$H(\operatorname{div}; \Omega_F) := \{ \tau \in [L^2(\Omega_F)]^{2 \times 2}; \operatorname{div} \tau \in [L^2(\Omega_F)]^2 \}$$

with norms

$$\begin{aligned} \|\tau\|_{[L^2(\Omega_F)]^{2 \times 2}} &:= \left(\sum_{i,j=1}^2 \|\tau_{ij}\|_{L^2(\Omega_F)}^2 \right)^{1/2} \\ \|\tau\|_{H(\operatorname{div}; \Omega_F)} &:= (\|\tau\|_{[L^2(\Omega_F)]^{2 \times 2}}^2 + \|\operatorname{div} \tau\|_{[L^2(\Omega_F)]^2}^2)^{1/2} \end{aligned}$$

and inner products

$$\int_{\Omega_F} \mathbf{s} : \boldsymbol{\tau} \, dx \quad \text{and} \quad \int_{\Omega_F} \mathbf{s} : \boldsymbol{\tau} \, dx + \int_{\Omega_F} \operatorname{div} \mathbf{s} \cdot \operatorname{div} \boldsymbol{\tau} \, dx$$

Here, the symbol \cdot denotes the product

$$\mathbf{s} : \boldsymbol{\tau} := \sum_{i,j=1}^2 s_{ij} \tau_{ij} \quad \text{for } \mathbf{s}, \boldsymbol{\tau} \in [L^2(\Omega_F)]^{2 \times 2}$$

and

$$\operatorname{div} \boldsymbol{\tau} := \begin{bmatrix} \operatorname{div}(\tau_{11} \, \tau_{12}) \\ \operatorname{div}(\tau_{21} \, \tau_{22}) \end{bmatrix} \quad \text{for } \boldsymbol{\tau} = \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix}$$

We also define

$$\mathbf{H}_0(\operatorname{div}; \Omega_F) := \{ \boldsymbol{\tau} \in \mathbf{H}(\operatorname{div}; \Omega_F); (\boldsymbol{\tau} \cdot \mathbf{n})|_{\Gamma_F} = 0 \}$$

Now, we follow Brink, Carstensen and Stein (1996) in deriving a weak formulation of the problem (169). On Ω_F , we consider \mathbf{u} and $\boldsymbol{\sigma}$ as independent unknowns. Testing (169b) with $\boldsymbol{\tau} \in \mathbf{H}_0(\operatorname{div}; \Omega_F)$, this gives

$$\int_{\Omega_F} \boldsymbol{\tau} : \mathbf{C}^{-1} \boldsymbol{\sigma} \, dx = \int_{\Omega_F} \boldsymbol{\tau} : \mathbf{e}(\mathbf{u}) \, dx$$

Integrating by parts and using $\mathbf{u} = \mathbf{g}$ on Γ_D , we obtain

$$\begin{aligned} \int_{\Omega_F} \boldsymbol{\tau} : \mathbf{C}^{-1} \boldsymbol{\sigma} \, dx + \int_{\Omega_F} \mathbf{u} \cdot \operatorname{div} \boldsymbol{\tau} \, dx \\ + \int_{\Omega_F} \boldsymbol{\tau} : \boldsymbol{\gamma} \, dx - \langle \boldsymbol{\phi}, \boldsymbol{\tau} \cdot \mathbf{v} \rangle_{\Gamma} \\ = \langle \mathbf{g}, \boldsymbol{\tau} \cdot \mathbf{v} \rangle_{\Gamma_D} \quad \forall \boldsymbol{\tau} \in \mathbf{H}_0(\operatorname{div}; \Omega_F) \end{aligned} \quad (170)$$

where $\boldsymbol{\gamma} := 1/2(\nabla \mathbf{u} - (\nabla \mathbf{u})^T)$ and

$$\boldsymbol{\phi} := \mathbf{u}|_{\Gamma} \in [H^{1/2}(\Gamma)]^2 \quad (171)$$

are introduced as further unknowns. Here, $\langle \cdot, \cdot \rangle_{\Gamma}$ stands for the duality pairing between $[H^{1/2}(\Gamma)]^2$ and $[H^{-1/2}(\Gamma)]^2$, whereas $\langle \cdot, \cdot \rangle_{\Gamma_D}$ denotes the duality between $[H^{1/2}(\Gamma_D)]^2$ and its dual $[H^{-1/2}(\Gamma_D)]^2$. We note that $\boldsymbol{\gamma}$ represents rotations and lies in the space

$$\begin{aligned} \mathbf{H}_0 &:= \{ \boldsymbol{\tau} \in [L^2(\Omega_F)]^{2 \times 2}; \boldsymbol{\tau} + \boldsymbol{\tau}^T = 0 \} \\ &= \left\{ \begin{bmatrix} 0 & \boldsymbol{\tau} \\ -\boldsymbol{\tau} & 0 \end{bmatrix}; \boldsymbol{\tau} \in L^2(\Omega_F) \right\} \end{aligned}$$

Further, testing (169a), we find that there holds

$$\int_{\Omega_F} \mathbf{v} \cdot \operatorname{div} \boldsymbol{\sigma} \, dx = - \int_{\Omega_F} \mathbf{f} \cdot \mathbf{v} \, dx \quad \forall \mathbf{v} \in [L^2(\Omega_F)]^2 \quad (172)$$

The problem of linear elasticity within Ω_B is dealt with by boundary integral operators on Γ . Denoting by \mathbf{V} , \mathbf{K} , \mathbf{K}' , and \mathbf{W} the integral operators of the single-layer, double-layer, adjoint of the double layer and hypersingular potentials respectively, and using the well-known jump conditions (cf. e.g. Hsiao, Stephan and Wendland, 1991), we obtain

$$2\boldsymbol{\phi} = (\mathbf{I} + \mathbf{K})\boldsymbol{\phi} - \mathbf{V}(\boldsymbol{\sigma} \cdot \mathbf{v}) \quad \text{on } \Gamma \quad (173)$$

and

$$\mathbf{W}\boldsymbol{\phi} + (\mathbf{I} + \mathbf{K}')(\boldsymbol{\sigma} \cdot \mathbf{v}) = 0 \quad \text{on } \Gamma \quad (174)$$

Here, $\boldsymbol{\phi} = \mathbf{u}|_{\Gamma}$ (cf. (171)) and \mathbf{I} is the identity operator on the corresponding spaces.

Eventually, we substitute (173) into (170) and we test (174) with the functions $\boldsymbol{\psi} \in [H^{1/2}(\Gamma)]^2$. Then, collecting (170), (172), and (174) and requiring the symmetry of $\boldsymbol{\sigma}$ weakly by

$$\int_{\Omega_F} \boldsymbol{\sigma} : \boldsymbol{\delta} \, dx = 0 \quad \forall \boldsymbol{\delta} \in \mathbf{H}_0$$

we arrive at the following variational formulation of the boundary value problem (169): find $(\boldsymbol{\sigma}, \boldsymbol{\phi}, \mathbf{u}, \boldsymbol{\gamma}) \in \mathbf{H}_0(\operatorname{div}; \Omega_F) \times [H^{1/2}(\Gamma)]^2 \times H_0$ such that

$$\begin{aligned} a(\boldsymbol{\sigma}, \boldsymbol{\phi}; \boldsymbol{\tau}, \boldsymbol{\psi}) + b(\boldsymbol{\tau}; \mathbf{u}, \boldsymbol{\gamma}) &= 2\langle \mathbf{g}, \boldsymbol{\tau} \cdot \mathbf{v} \rangle_{\Gamma_D} \\ b(\boldsymbol{\sigma}; \mathbf{v}, \boldsymbol{\delta}) &= -2 \int_{\Omega_F} \mathbf{f} \cdot \mathbf{v} \, dx \end{aligned} \quad (175)$$

for all $(\boldsymbol{\tau}, \boldsymbol{\psi}, \mathbf{v}, \boldsymbol{\delta}) \in \mathbf{H}_0(\operatorname{div}; \Omega_F) \times [H^{1/2}(\Gamma)]^2 \times [L^2(\Omega_F)]^2 \times H_0$. Here,

$$\begin{aligned} a(\boldsymbol{\sigma}, \boldsymbol{\phi}; \boldsymbol{\tau}, \boldsymbol{\psi}) &= 2a_F(\boldsymbol{\sigma}, \boldsymbol{\tau}) + a_B(\boldsymbol{\sigma}, \boldsymbol{\phi}; \boldsymbol{\tau}) \\ &\quad - \langle \boldsymbol{\psi}, \mathbf{W}\boldsymbol{\phi} \rangle_{\Gamma} - \langle \boldsymbol{\psi}, (\mathbf{I} + \mathbf{K}')(\boldsymbol{\sigma} \cdot \mathbf{v}) \rangle_{\Gamma} \end{aligned}$$

with

$$\begin{aligned} a_F(\boldsymbol{\sigma}, \boldsymbol{\tau}) &:= \int_{\Omega_F} \boldsymbol{\tau} : \mathbf{C}^{-1} \boldsymbol{\sigma} \, dx \\ a_B(\boldsymbol{\sigma}, \boldsymbol{\phi}; \boldsymbol{\tau}) &:= \langle \boldsymbol{\tau} \cdot \mathbf{v}, \mathbf{V}(\boldsymbol{\sigma} \cdot \mathbf{v}) \rangle_{\Gamma} - \langle \boldsymbol{\tau} \cdot \mathbf{v}, (\mathbf{I} + \mathbf{K})\boldsymbol{\phi} \rangle_{\Gamma} \end{aligned}$$

and

$$b(\boldsymbol{\sigma}; \mathbf{v}, \boldsymbol{\delta}) = 2 \int_{\Omega_F} \mathbf{v} \cdot \operatorname{div} \boldsymbol{\sigma} \, dx + 2 \int_{\Omega_F} \boldsymbol{\sigma} : \boldsymbol{\delta} \, dx$$

Brink, Carstensen and Stein (1996) proved unique solvability of this weak formulation.

Theorem 20. For every $\mathbf{f} \in [L^2(\Omega_F)]^2$ and $\mathbf{g} \in [H^{1/2}(\Gamma_D)]^2$, the saddle-point problem (175) has a unique solution satisfying

$$\begin{aligned} \|\boldsymbol{\sigma}\|_{\mathbf{H}(\operatorname{div}; \Omega_F)} + \|\boldsymbol{\phi}\|_{[H^{1/2}(\Gamma)]^2} + \|\mathbf{u}\|_{[L^2(\Omega_F)]^2} \\ + \|\boldsymbol{\gamma}\|_{[L^2(\Omega_F)]^{2 \times 2}} \leq C \{ \|\mathbf{f}\|_{[L^2(\Omega_F)]^2} + \|\mathbf{g}\|_{[H^{1/2}(\Gamma_D)]^2} \} \end{aligned}$$

with a positive constant C , which is independent of \mathbf{f} and \mathbf{g} .

Let $\bar{\Omega}_F = \cup \{ \bar{T}; T \in \mathcal{T}_h \}$ be a partitioning of Ω_F . The elements T are triangles or quadrilaterals. For $T \neq T'$, $\bar{T} \cap \bar{T}'$ is either empty or a common vertex or edge.

To define an implicit residual-error estimator, we follow the strategy from Section 2.2.3 and use an elliptic, continuous, symmetric bilinear form $\hat{a}(\cdot, \cdot)$ on $\mathbf{H}_0(\operatorname{div}; \Omega_F) \times [H^{1/2}(\Gamma)]^2$. Here, $[H^{1/2}(\Gamma)]^2$ is the quotient space $[H^{1/2}(\Gamma)]^2 / \ker(\mathbf{e}|_{\Gamma})$ that eliminates the rigid body motions $\ker(\mathbf{e}|_{\Gamma})$. For simplicity, we take the bilinear forms

$$\begin{aligned} \hat{a}_F(\boldsymbol{\sigma}, \boldsymbol{\tau}) &:= \int_{\Omega_F} \boldsymbol{\sigma} : \boldsymbol{\tau} \, dx + \int_{\Omega_F} \operatorname{div} \boldsymbol{\sigma} \cdot \operatorname{div} \boldsymbol{\tau} \, dx \\ &\quad \text{on } \mathbf{H}_0(\operatorname{div}; \Omega_F) \times \mathbf{H}_0(\operatorname{div}; \Omega_F) \end{aligned}$$

and

$$\hat{a}_B(\boldsymbol{\phi}, \boldsymbol{\psi}) := \langle \boldsymbol{\psi}, \mathbf{W}\boldsymbol{\phi} \rangle_{\Gamma} \quad \text{on } [H^{1/2}(\Gamma)]^2 \times [H^{1/2}(\Gamma)]^2$$

and define

$$\hat{a}(\boldsymbol{\sigma}, \boldsymbol{\phi}; \boldsymbol{\tau}, \boldsymbol{\psi}) := 2\hat{a}_F(\boldsymbol{\sigma}, \boldsymbol{\tau}) + \hat{a}_B(\boldsymbol{\phi}, \boldsymbol{\psi}) \quad (176)$$

Note that \hat{a} is an inner product on $\mathbf{H}_0(\operatorname{div}; \Omega_F) \times [H^{1/2}(\Gamma)]^2$. The restriction of \hat{a}_F to an element $T \in \mathcal{T}_h$ is denoted by $\hat{a}_{F,T}$, that is,

$$\hat{a}_F(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \sum_{T \in \mathcal{T}_h} \hat{a}_{F,T}(\boldsymbol{\sigma}|_T, \boldsymbol{\tau}|_T) \quad \text{for all } \boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbf{H}_0(\operatorname{div}; \Omega_F)$$

Now, let us assume that we have an approximate solution $(\boldsymbol{\sigma}_h, \boldsymbol{\phi}_h, \mathbf{u}_h, \boldsymbol{\gamma}_h) \in \mathbf{H}_0(\operatorname{div}; \Omega_F) \times [H^{1/2}(\Gamma)]^2 \times [L^2(\Omega_F)]^2 \times H_0$ to (175). In practice, this will be a coupled finite element/BE solution obtained by restricting (175) to some discrete subspaces.

For the error estimator, the following local problems are defined: for $T \in \mathcal{T}_h$, find

$$\begin{aligned} \boldsymbol{\sigma}_T \in \mathbf{H}_0(\operatorname{div}; T) &:= \{ \boldsymbol{\tau} \in [L^2(T)]^{2 \times 2}; \\ &\quad \operatorname{div} \boldsymbol{\tau} \in [L^2(T)]^2, (\boldsymbol{\tau} \cdot \mathbf{v})|_{\partial T \cap \Gamma_F} = 0 \} \end{aligned}$$

such that

$$\hat{a}_{F,T}(\boldsymbol{\sigma}_T, \boldsymbol{\tau}) = -F_T(\boldsymbol{\tau}) \quad \text{for all } \boldsymbol{\tau} \in \mathbf{H}_0(\operatorname{div}; T) \quad (177)$$

where

$$\begin{aligned} F_T(\boldsymbol{\tau}) &:= a_{F,T}(\boldsymbol{\sigma}_h, \boldsymbol{\tau}) + a_{B,T}(\boldsymbol{\sigma}_h, \boldsymbol{\phi}_h; \boldsymbol{\tau}) \\ &\quad + b_T(\boldsymbol{\tau}; \mathbf{u}_h, \boldsymbol{\gamma}_h) - 2 \int_{\partial T \cap \Gamma} \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} \cdot \mathbf{n}) \, ds \end{aligned}$$

(\mathbf{n} is the outward unit normal of T) with

$$\begin{aligned} a_{F,T}(\boldsymbol{\sigma}_h, \boldsymbol{\tau}) &:= \int_T \boldsymbol{\tau} : \mathbf{C}^{-1} \boldsymbol{\sigma}_h \, dx \\ a_{B,T}(\boldsymbol{\sigma}_h, \boldsymbol{\phi}_h; \boldsymbol{\tau}) &:= \langle \boldsymbol{\tau} \cdot \mathbf{v}, \mathbf{V}(\boldsymbol{\sigma}_h \cdot \mathbf{v}) \rangle_{\partial T \cap \Gamma} \\ &\quad - \langle \boldsymbol{\tau} \cdot \mathbf{v}, (\mathbf{I} + \mathbf{K})\boldsymbol{\phi}_h \rangle_{\partial T \cap \Gamma} \\ b_T(\boldsymbol{\tau}; \mathbf{u}_h, \boldsymbol{\gamma}_h) &:= 2 \int_T \mathbf{u}_h \cdot \operatorname{div} \boldsymbol{\tau} \, dx + 2 \int_T \boldsymbol{\tau} : \boldsymbol{\gamma}_h \, dx \end{aligned}$$

Here, $\boldsymbol{\lambda} \in [H^{1/2}(\cup_{T \in \mathcal{T}_h} \partial T \setminus \Gamma)]^2$ is arbitrary on the element boundaries interior to Ω_F and on Γ_N . We require that $\boldsymbol{\lambda} = \mathbf{g}$ on Γ_D . Note that if $\boldsymbol{\lambda}|_{\partial T \cap \Gamma_F \cup \Gamma} = \mathbf{u}|_{\partial T \cap \Gamma_F \cup \Gamma}$ for $T \in \mathcal{T}_h$ (which in particular implies that $\boldsymbol{\lambda}|_{\Gamma_D} = \mathbf{g}$), then the solution $\boldsymbol{\sigma}_T$ of (177) converges to 0 in $\mathbf{H}_0(\operatorname{div}; T)$ if $(\boldsymbol{\sigma}_h, \boldsymbol{\phi}_h, \mathbf{u}_h, \boldsymbol{\gamma}_h)$ converges to $(\boldsymbol{\sigma}, \boldsymbol{\phi}, \mathbf{u}, \boldsymbol{\gamma})$. The local solution $\boldsymbol{\sigma}_T$ can be considered as a projection of the error $(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\phi} - \boldsymbol{\phi}_h, \mathbf{u} - \mathbf{u}_h, \boldsymbol{\gamma} - \boldsymbol{\gamma}_h)$ onto the local space $\mathbf{H}_0(\operatorname{div}; T)$. We expect that a good approximation of $\boldsymbol{\lambda}$ to \mathbf{u} on interior element edges improves the efficiency of our error estimator.

In Gatica, Heuer and Stephan (2001), we prove the following a posteriori error estimate based on the local problems (177) yielding reliability. The proof is a modification of the proof of Theorem 7.

Theorem 21. Let $\boldsymbol{\lambda} \in [H^{1/2}(\cup_{T \in \mathcal{T}_h} \partial T \setminus \Gamma)]^2$ with $\boldsymbol{\lambda}|_{\Gamma_D} = \mathbf{g}$. Further, define for any $T \in \mathcal{T}_h$ the function $\boldsymbol{\sigma}_T \in \mathbf{H}_0(\operatorname{div}; T)$ by the local problem (177). Then, there holds the a posteriori error estimate

$$\begin{aligned} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{H}(\operatorname{div}; \Omega_F)} + \|\boldsymbol{\phi} - \boldsymbol{\phi}_h\|_{[H^{1/2}(\Gamma)]^2} \\ + \|\mathbf{u} - \mathbf{u}_h\|_{[L^2(\Omega_F)]^2} + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_{[L^2(\Omega_F)]^{2 \times 2}} \\ \leq C \left\{ \left[\sum_{T \in \mathcal{T}_h} \hat{a}_{F,T}(\boldsymbol{\sigma}_T, \boldsymbol{\sigma}_T) + R_T^2 \right]^{1/2} \right. \\ \left. + \|\mathbf{f} + \operatorname{div} \boldsymbol{\sigma}_h\|_{[L^2(\Omega_F)]^2} + \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^*\|_{[L^2(\Omega_F)]^{2 \times 2}} \right\} \end{aligned}$$

where C depends on the norm of

$$\mathbf{W}^{-1}: [H^{-1/2}(\Gamma)]^2 \rightarrow [H^{1/2}(\Gamma)]^2 / \ker \mathbf{W}$$

and on the inf-sup constant β of the bilinear form of the saddle-point problem (175), but is independent of h . Here,

R_T is the norm of the residual for the boundary integral equation (174), that is,

$$R_T := \|W\phi_h + (I + K')(\sigma_h \cdot \nu)\|_{[H^{-1/2}(\Gamma)]^2}$$

and the kernel of W consists of the rigid body motions, that is, $[H^{1/2}(\Gamma)]^2 / \ker W = [H^{1/2}(\Gamma)]_0^2$.

Note that the above error estimate does not make use of any special FE or BE spaces. Here, the residual term is given in a negative-order Sobolev norm. In practical applications, in which a certain BE subspace is used, this norm can be estimated by weighted local L^2 -norms.

7 CONCLUDING REMARKS

This chapter gives an overview of boundary element and finite element coupling procedures for elliptic interface problems. Special emphasis is given to the derivation of a posteriori error estimates. Owing to the limitation of the chapter length, we have omitted to describe corresponding adaptive algorithms that can be found in the given literature. The FE/BE coupling method is still a rapidly developing field, and this chapter can only describe some fundamental concepts. Further issues, which are not discussed above, are, for example, the following topics.

In BEM implementations for engineering applications, the integral equation is often discretized by collocation (see Brebbia, Telles and Wrobel, 1984). The coupling of FEM and Galerkin BEM was analyzed by Wendland (1988), and Galerkin BEM by Wendland (1990). In Wendland (1988), the mesh refinement requires the condition $k = o(h)$, with $k(h)$ denoting the sizes of the BE (FE) mesh. In Wendland (1990) and in Brink and Stephan (1996), this condition is weakened to $k \leq \beta \cdot h$, $\beta \in \mathbb{R}$, and convergence is shown on the basis of the ideas from Brezzi and Johnson (1979). To establish convergence, the essential observation is that, given an FEM mesh, it suffices to solve the boundary integral equations accurately enough. In Brink and Stephan (1996), we show that in the energy norm the coupled method converges with optimal order.

A hybrid coupled FE/BE method for linear elasticity problems is presented in Hsiao (1990) and Hsiao, Schnack and Wendland (1999, 2000). In this hybrid method, in addition to traditional FE, Trefftz elements are considered. These are modeled with boundary potentials supported by the individual element boundaries, the so-called macroelements. Collocation as well as Galerkin methods are used for the coupling between the FE and macroelements. The coupling is realized via mortar elements on the coupling boundary. Here, different discretizations on Ω_1 and Ω_2 are coupled via weak formulations and the use of mortar or

Lagrange spaces on the coupling boundaries (cf. Steinbach, 2003).

For a parabolic-elliptic interface problem with the heat equation in the interior domain and the Laplace equation in the exterior domain, modeling two-dimensional eddy currents in electrodynamics, Costabel, Ervin, Stephan (1990) introduce a full discretization of the problem by symmetric coupling of FE and BE. For the discretization in time, the Crank-Nicolson method is proposed there. In Mund and Stephan (1997), we use the discontinuous Galerkin method (with piecewise linear test and trial functions), which allows space and time steps to be variable in time. On the basis of an a posteriori error estimate of residual type, we present a reliable adaptive algorithm for choosing local mesh sizes in space and time. The linear systems of equations obtained by the above mentioned discretization in space and time are symmetric and indefinite; they are solved by the HMCN method (see Chandra, Eisenstat and Schultz, 1977), a stable version of MINRES.

Symmetric and nonsymmetric coupling FE/BE procedures for solving the heterogeneous Maxwell equations in $\mathbb{R}^3 \setminus \Omega_\infty$ with a Leontovich boundary condition on Γ_∞ are given in Ammari and Nédélec (2000). In this paper, the authors consider the time-harmonic electromagnetic scattering by a bounded dielectric material Ω surrounding a lossy highly conductive body Ω_∞ .

In Teltscher, Maischak and Stephan (2003), we present a symmetric FE/BE coupling method for the time-harmonic eddy-current problem (in electro magnetics) for low frequencies.

The symmetric coupling method (see also Kuhn and Steinbach (2002) and Hiptmair (2002)) is based on a weak formulation of the vector Helmholtz equation for the electrical field u in the polyhedral domain Ω and uses a weak formulation of the integral equations for $u|_\Gamma$, the trace of u on Γ , and for λ , the twisted tangential trace of the magnetic field on Γ , the boundary of the conductor. The resulting system is strongly elliptic and symmetric and yields quasi-optimal convergence for any conforming Galerkin coupling method. We present a posteriori error estimates of residual and hierarchical type. Numerical results with lowest-order Nédélec elements on hexahedra as FE and lowest-order RT elements (with vanishing surface divergence) as BE underline our theory.

In Brink and Stephan (2001), the FE/BE coupling is analyzed, in which, in the FEM domain, we assume an incompressible elastic material and use a Stokes-type mixed FEM; linear elasticity is considered in the BEM domain. For rubberlike materials, the incompressibility has to be taken into account, which requires mixed FE. The stress (which is often the physical quantity of maximum interest) cannot be determined merely from the displacement, if the

material is incompressible. Additional unknowns have to be introduced. In Brink and Stephan (2001), we employ a primal-mixed finite element method with the pressure as the secondary unknown.

Alternatively, one may use dual-mixed methods, in which, in elasticity, the stress tensor is the primary unknown. A coupling of BEM and dual-mixed FEM was proposed by Brink, Carstensen and Stein (1996), in which linear elasticity was assumed in the FEM domain.

For problems with nonlinearities, a so-called dual-dual formulation can be applied, since it avoids to invert the elasticity tensor C directly (see Gatica and Heuer, 2000). An a posteriori error estimate for the pure FEM (no coupling with BEM), based on the dual-dual formulation, is given in Barrientos, Gatica and Stephan (2002). Suitable preconditioned solvers for the dual-dual coupling can be found in Gatica and Heuer (2002).

An alternative procedure to the FE/BE coupling for exterior nonlinear-linear transmission problems consists of employing DtN mappings instead of BEM (cf. Givoli, 1992). This means that one first introduces a sufficiently large circle Γ (in \mathbb{R}^2), or a sphere (in \mathbb{R}^3), such that the linear domain is divided into a bounded annular region and an unbounded one. Then, one derives an explicit formula for the Neumann data on Γ in terms of the Dirichlet data on the same curve, which is known as the DtN mapping. This has been done for several elliptic operators, including the Lamé system for elasticity, and acoustic problems, using Fourier-type series developments (see MacCamy and Marin, 1980; Gatica, 1997; Barrenechea, Gatica and Hsiao, 1998; Barrientos, Gatica and Maischak, 2002; and Gatica, Gatica and Stephan, 2003).

REFERENCES

- Ammari H and Nédélec JC. Coupling integral equation methods and finite volume elements for the resolution of the Leontovich boundary value problem for the time-harmonic Maxwell equations in three dimensional heterogeneous media. In *Mathematical Aspects of Boundary Element Methods*. CRC Research Notes in Mathematics 414, Bonnet M, Sändig AM and Wendland WL (eds). Chapman & Hall: Boca Raton, 2000; 11–22.
- Ashby S, Mantouffou T and Saylor PE. A taxonomy for conjugate gradient methods. *SIAM J. Numer. Anal.* 1990; 27(6):1542–1568.
- Babuska I, Craig A, Mandel J and Pitkäranta J. Efficient preconditioning for the p-version finite element method in two dimensions. *SIAM J. Numer. Anal.* 1991; 28(3):624–661.
- Bank RE and Smith R. A posteriori error estimates based on hierarchical bases. *SIAM J. Numer. Anal.* 1993; 30(4):921–935.

Bank RE and Weiser A. Some a posteriori error estimators for elliptic partial differential equations. *Math. Comp.* 1985; 44(170):283–301.

Barrenechea GR, Gatica GN and Hsiao GC. Weak solvability of interior transmission problems via mixed finite elements and Dirichlet-to-Neumann mappings. *J. Comput. Appl. Math.* 1998; 100:145–160.

Barrientos MA, Gatica G and Maischak M. A posteriori error estimates for linear exterior problems via mixed-FEM and DtN mappings. *Math. Modell. Numer. Anal.* 2002; 36:241–272.

Barrientos MA, Gatica G and Stephan EP. A mixed finite element method for nonlinear elasticity: two-fold saddle point approach and a posteriori error estimate. *Numer. Math.* 2002; 91(2):197–222.

Bettess P, Kelly DW and Zienkiewicz OC. The coupling of the finite element method and boundary solution procedures. *Int. J. Numer. Math. Eng.* 1977; 11:355–375.

Bielak J and MacCamy RC. An exterior interface problem in two-dimensional elastodynamics. *Q. Appl. Math.* 1983; 41(1):143–159.

Bramble J and Pasciak J. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comp.* 1988; 50(181):1–17.

Bramble J, Lazarov R and Pasciak J. A least-squares approach based on a discrete minus one inner product for first order systems. *Math. Comp.* 1997; 66(219):935–955.

Bramble J, Leyk Z and Pasciak J. The analysis of multigrid algorithms for pseudo-differential operators of order minus one. *Math. Comp.* 1994; 63(208):461–478.

Brebbia CA, Telles JCF and Wrobel LC. *Boundary Element Techniques*. Springer-Verlag: Berlin, 1984.

Brezzi F and Johnson C. On the coupling of boundary integral and finite element methods. *Calcolo* 1979; 16(2):189–201.

Brink U and Stephan EP. Convergence rates for the coupling of FEM and collocation BEM. *IMA J. Numer. Anal.* 1996; 16(1):93–110.

Brink U and Stephan EP. Implicit residual error estimators for the coupling of finite elements and boundary elements. *Math. Methods Appl. Sci.* 1999; 22(11):923–936.

Brink U and Stephan EP. Adaptive coupling of boundary elements and mixed finite elements for incompressible elasticity. *Numer. Methods Part. Differ. Equations* 2001; 17(1):79–92.

Brink U, Carstensen C and Stein E. Symmetric coupling of boundary elements and Raviart-Thomas type mixed finite elements in elastostatics. *Numer. Math.* 1996; 75(2):153–174.

Carstensen C. Interface problem in holonomic elastoplasticity. *Math. Methods Appl. Sci.* 1993; 16(11):819–835.

Carstensen C. Interface problems in viscoplasticity and plasticity. *SIAM J. Math. Anal.* 1994; 25(6):1468–1487.

Carstensen C. Coupling of FEM and BEM for interface problems in viscoplasticity and plasticity with hardening. *SIAM J. Numer. Anal.* 1996; 33(1):171–207.

Carstensen C. A posteriori error estimates for the symmetric coupling of finite elements and boundary elements. *Computing* 1996; 57(4):301–322.

- Carstensen C, Funken SA and Stephan EP. On the adaptive coupling of fem and bem in 2-d elasticity. *Numer. Math.* 1997; 77(2):187–221.
- Carstensen C and Gwinner J. FEM and BEM coupling for a nonlinear transmission problem with Signorini contact. *SIAM J. Numer. Anal.* 1997; 34(5):1845–1864.
- Carstensen C and Stephan EP. Adaptive coupling of boundary and finite element methods. *RAIRO Modell. Math. Anal. Numer.* 1995; 29(7):779–817.
- Carstensen C and Stephan EP. Coupling of FEM and BEM for a nonlinear interface problem: the h-p version. *Numer. Methods Part. Differ. Equations* 1995; 11(5):539–554.
- Carstensen C and Stephan EP. Interface problems in elastoviscoplasticity. *Q. Appl. Math.* 1995; 53(4):633–655.
- Carstensen C, Kühn M and Langer U. Fast parallel solvers for symmetric boundary element domain decomposition methods. *Numer. Math.* 1998; 79:321–347.
- Carstensen C, Zarrabi D and Stephan EP. On the h-adaptive coupling of FE and BE for viscoplastic and elasto-plastic interface problems. *J. Comput. Appl. Math.* 1996; 75(2):345–363.
- Chandra R, Eisenstat S and Schuit M. The modified conjugate residual method for partial differential equations. In *Advances in Computer Methods for Partial Differential Equations II*, Vichnevsky R (ed.). IMACS: New Brunswick, 1977; 13–19.
- Costabel M. Symmetric methods for the coupling of finite elements and boundary elements. In *Boundary Elements IX*, Brebbia CA and Wendland WL (eds), Springer, 1987; 411–420.
- Costabel M. Boundary integral operators on Lipschitz domains: elementary results. *SIAM J. Math. Anal.* 1988; 19(3):613–626.
- Costabel M. A symmetric method for the coupling of finite elements and boundary elements. In *The Mathematics of Finite Elements and Applications, VI, MAFELAP 1987*, Whiteman J (ed.), Academic Press: London, 1988; 281–288.
- Costabel M and Stephan EP. Coupling of finite elements and boundary elements for transmission problems of elastic waves in \mathbb{R}^3 . In *Advanced Boundary Element Methods, IUTAM Symposium 1987*, Cruse TA (ed.), Springer: Berlin, San Antonio, 1988; 117–124.
- Costabel M and Stephan EP. Coupling of finite elements and boundary elements for inhomogeneous transmission problems in \mathbb{R}^3 . In *The Mathematics of Finite Elements and Applications, VI, MAFELAP 1987*, Whiteman J (ed.), Academic Press: London, 1988; 289–296.
- Costabel M and Stephan EP. Coupling of finite and boundary element methods for an elastoplastic interface problem. *SIAM J. Numer. Anal.* 1990; 27(5):1212–1226.
- Costabel M, Ervin VJ and Stephan EP. Symmetric coupling of finite elements and boundary elements for a parabolic-elliptic interface problem. *Q. Appl. Math.* 1990; 48(2):265–279.
- Costabel M, Ervin VJ and Stephan EP. Experimental convergence rates for various couplings of boundary and finite elements. *Math. Comput. Modell.* 1991; 15(3–5):93–102.
- Eck C, Schulz H, Steinbach O and Wendland WL. An adaptive boundary element method for contact problems. In *Error-Controlled Adaptive Finite Elements in Solid Mechanics*, Stein E (ed.), John Wiley & Sons: Chichester, 2003; 181–209.
- Ekeland I and Temam R. Analyse convexe et problèmes variationnels. *Études Mathématiques*. Dunod, Gauthier – Villars: Paris – Bruxelles – Montreal, 1974; 1–340.
- Ervin VJ, Heuer N and Stephan EP. On the h-p version of the boundary element method for symm's integral equation on polygons. *Comput. Methods Appl. Mech. Eng.* 1993; 110(1–2):25–38.
- Funken SA and Stephan EP. *Fast Solvers with Multigrid Preconditioners for Linear Fem-bem Coupling*. Preprint, University of Hannover: Hannover, 2001.
- Gaier D. Integralgleichungen erster Art und konforme Abbildungen. *Math. Zeitschr.* 1976; 147:113–139.
- Gatica G. Combination of mixed finite element and Dirichlet-to-Neumann methods in non-linear plane elasticity. *Appl. Math. Lett.* 1997; 10(6):29–35.
- Gatica G and Heuer N. A FEM-DIN formulation for a non-linear exterior problem in incompressible elasticity. *Math. Methods Appl. Sci.* 2003; 26(2):151–170.
- Gatica G and Heuer N. A dual-dual formulation for the coupling of mixed-FEM and BEM in hyperelasticity. *SIAM J. Numer. Anal.* 2000; 38(2):380–400.
- Gatica G and Heuer N. A preconditioned MINRES method for the coupling of mixed-FEM and BEM in some nonlinear problems. *SIAM J. Sci. Comput.* 2002; 24(2):572–596.
- Gatica G and Hsiao GC. The coupling of boundary element and finite element methods for a nonlinear exterior boundary value problem. *Z. Anal. Angew.* 1989; 8(4):377–387.
- Gatica G and Hsiao GC. On a class of variational formulations for some non-linear interface problems. *Rendiconti Math.* 1990; 10:681–715.
- Gatica G and Hsiao GC. *Boundary Field Equation Methods for a Class of Nonlinear Problems*. Longman: Harlow, 1995.
- Gatica G, Harbrecht H and Schneider R. *Least Squares Methods for the Coupling of Fem and Bem*. Preprint, Universidad de Concepción, 2001.
- Gatica G, Heuer N and Stephan EP. An implicit-explicit residual error estimator for the coupling of dual-mixed finite elements and boundary elements in elastostatics. *Math. Methods Appl. Sci.* 2001; 24(3):179–191.
- Gatica G, Maishchak M and Stephan EP. *On the Coupling of Mixed Finite Element and Boundary Element Methods for a Transmission Problem with Signorini Contact*. Preprint, Universidad de Concepción, 2003 (submitted).
- Girault V and Raviart PA. *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*. Springer-Verlag, 1986.
- Givoli D. *Numerical Methods for Problems in Infinite Domains*. Elsevier: Amsterdam, 1992.
- Guo BQ and Stephan EP. The h-p version of the coupling of finite element and boundary element methods for transmission problems in polyhedral domains. *Numer. Math.* 1998; 80(1):87–107.
- Gwinner J and Stephan EP. A boundary element procedure for contact problems in linear elastostatics. *RAIRO Math. Modell. Numer. Anal.* 1993; 27(4):457–480.
- Hahne M, Stephan EP and Thies W. Fast solvers for coupled fem – bem equations I. In *Fast Solvers for Flow Problems*, Vol. 49 Notes on Numerical Fluid Mechanics, Hackbusch W and Wittum G (eds), Vieweg: Braunschweig, 1995; 121–130.
- Hahne M, Maishchak M, Stephan EP and Wathen A. Efficient preconditioners for coupled fem – bem equations. *Numer. Methods Part. Differ. Equations* to appear.
- Han H. A new class of variational formulations for the coupling of finite and boundary element methods. *J. Comput. Math.* 1990; 8:223–232.
- Heuer N and Stephan EP. Coupling of the finite element and boundary element method. The h-p version. *ZAMM* 1991; 71(6):T584–T586.
- Heuer N and Stephan EP. Preconditioners for the p-version of the Galerkin method for a coupled finite element/boundary element system. *Numer. Methods Part. Differ. Equations* 1998; 14(1):47–61.
- Heuer N and Stephan EP. An additive Schwarz method for the h-p version of the boundary element method for hypersingular integral equations in \mathbb{R}^3 . *IMA J. Numer. Anal.* 2001; 21(1):263–283.
- Heuer N, Maishchak M and Stephan EP. Preconditioned minimum residual iteration for the h-p version of the coupled fem-bem with quasinormed meshes. *Numer. Lin. Alg. Appl.* 1999; 6(6):435–456.
- Heuer N, Stephan EP and Tran T. Multilevel additive Schwarz method for the h-p version of the Galerkin boundary element method. *Math. Comp.* 1998; 67(222):501–518.
- Hiptmair R. Symmetric coupling for eddy current problems. *SIAM J. Numer. Anal.* 2002; 40(1):41–65.
- Hsiao GC. The coupling of boundary and finite element methods. *Z. Angew. Math. Mech.* 1990; 70(6):T493–T505.
- Hsiao GC. Some recent developments on the coupling of finite element and boundary element methods. *Rend. Sem. Mat. Univ. Torino. Fascicolo Speciale* 1991; *Numer. Methods* 1992; 95–111.
- Hsiao GC and Han H. The boundary element method for a contact problem. In *Theory and Applications of Boundary Elements*, Du Q and Tanaka M (eds), Tsinghua University Press: Beijing, 1988; 33–38.
- Hsiao GC and Wendland WL. A finite element method for some integral equations of the first kind. *J. Math. Anal. Appl.* 1977; 58:449–481.
- Hsiao GC, Schnack E and Wendland WL. A hybrid coupled finite-boundary element method in elasticity. *Comput. Methods Appl. Mech. Eng.* 1999; 173(3–4):287–316.
- Hsiao GC, Schnack E and Wendland WL. Hybrid coupled finite-boundary element methods for elliptic systems of second order. *Comput. Methods Appl. Mech. Eng.* 2000; 190:431–485.
- Hsiao GC, Steinbach O and Wendland WL. Domain decomposition methods via boundary integral equations. *J. Comput. Appl. Math.* 2000; 125:523–539.
- Hsiao GC, Stephan EP and Wendland WL. On the Dirichlet problem in elasticity for a domain exterior to an arc. *J. Comput. Appl. Math.* 1991; 34(1):1–19.
- Johnson C and Nedelec JC. On the coupling of boundary integral and finite element methods. *Math. Comp.* 1980; 35(152):1063–1079.
- Krebs A, Maishchak M and Stephan EP. Adaptive FEM-BEM Coupling with a Schur Complement Error Indicator. 2001, submitted. <http://ftp.ifam.uni-hannover.de/pub/preprints/ifam45.ps.Z>.
- Kuhn M and Steinbach O. Symmetric coupling of finite and boundary elements for exterior magnetic problems. *Math. Methods Appl. Sci.* 2002; 25:357–371.
- Langer U. Parallel iterative solution of symmetric coupled fe/be equations via domain decomposition. *Contemp. Math.* 1994; 157:335–344.
- Lebedev V. Iterative methods for solving operator equations with a spectrum contained in several intervals. *USSR Comput. Math. Math. Phys.* 1969; 9:17–24.
- MacCamy RC and Martin SP. A finite element method for exterior interface problems. *Int. J. Math. Math. Sci.* 1980; 3:311–350.
- Maishchak M. *FEM/BEM Coupling Methods for Signorini-type Interface Problems – Error Analysis Adaptivity, Preconditioners*. Habilitationsschrift, University of Hannover, 2001.
- Maishchak M. *Manual of the Software Package Mairprogs*. Technical report ifam48, Institute for applied mathematics, University of Hannover, Hannover, 2003. <http://ftp.ifam.uni-hannover.de/pub/preprints/ifam48.ps.Z>.
- Maishchak M and Stephan EP. A least squares coupling method with finite elements and boundary elements for transmission problems. *Comput. Math. Appl.* to appear.
- Maishchak M, Stephan EP and Tran T. Domain decomposition for integral equations of the first kind: numerical results. *Appl. Anal.* 1996; 63(1–2):111–132.
- Maishchak M, Stephan EP and Tran T. Multiplicative Schwarz algorithms for the Galerkin boundary element method. *SIAM J. Numer. Anal.* 2000; 38(4):1243–1268.
- Maitre JF and Pourquieu O. Condition number and diagonal preconditioning: comparison of the p-version and the spectral element methods. *Numer. Math.* 1996; 74(1):69–84.
- Maue AW. Zur Formulierung eines allgemeinen Bewegungsproblems durch eine Integralgleichung. *Z. Phys.* 1949; 126:601–618.
- Meddahi S, Valdes J, Menendez O and Perez P. On the coupling of boundary integral and mixed finite element methods. *J. Comput. Appl. Math.* 1996; 69(1):113–124.
- Mund P and Stephan EP. Adaptive coupling and fast solution of FEM-BEM equations for parabolic-elliptic interface problems. *Math. Methods Appl. Sci.* 1997; 20(5):403–423.
- Mund P and Stephan EP. The preconditioned GMRES method for systems of coupled fem-bem equations. *Adv. Comput. Math.* 1998; 9(1–2):131–144.
- Mund P and Stephan EP. An adaptive two-level method for the coupling of nonlinear fem-bem equations. *SIAM J. Numer. Anal.* 1999; 36(4):1001–1021.
- Necas J. *Introduction to the Theory of Nonlinear Elliptic Equations*. Wiley-Interscience: Chichester, 1986.
- Necas J and Hlaváček I. *Mathematical Theory of Elastic and Elasto-Plastic Bodies*. Elsevier: Amsterdam, 1981.
- Nedelec JC. Integral equations with non integrable kernels. *Integral Equations Operator Theory* 1982; 5(4):562–572.
- O'Leary DP. A generalized conjugate gradient algorithm for solving a class of quadratic programming problems. *Numer. Lin. Alg. Appl.* 1980; 34:371–399.

- Paige CC and Saunders MA. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.* 1975; 12(4):617–629.
- Polizzotto C. A symmetric-definite BEM formulation for the elasto-plastic rate problem. In *Boundary Elements IX*, Brebbia CA, Wendland WL and Kuhn G (eds), vol. 2. Springer-Verlag: Berlin, 1987; 315–334.
- Sloan I and Spence A. The Galerkin method for integral equations of the first kind with logarithmic kernel: theory. *IMA J. Numer. Anal.* 1988; 8(1):105–122.
- Spann W. On the boundary element method for the Signorini problem of the Laplacian. *Numer. Math.* 1993; 65(3):337–356.
- Steinbach O. *Stability Estimates for Hybrid Coupled Domain Decomposition Methods*. Springer-Verlag: Berlin, 2003.
- Steinbach O and Wendland WL. Domain decomposition and preconditioning techniques in boundary element methods. *Boundary element topics* (Stuttgart, 1995). Springer: Berlin, 1997; 471–490.
- Steinbach O and Wendland WL. The construction of efficient preconditioners in the boundary element method. *Adv. Comput. Math.* 1998; 9:191–216.
- Steinbach O and Wendland WL. On C. Neumann's method for second order elliptic systems in domains with non-smooth boundaries. *J. Math. Anal. Appl.* 2001; 262:733–748.
- Stephan EP. Coupling of finite elements and boundary elements for some nonlinear interface problems. *Comput. Methods Appl. Mech. Engrg.* 1992; 101(1–3):61–72.
- Stephan EP and Wendland WL. Remarks to Galerkin and least squares methods with finite elements for general elliptic problems. *Manuscr. Geodaetica* 1976; 1:93–123.
- Stephan EP and Wendland WL. An augmented Galerkin procedure for the boundary integral method applied to two-dimensional screen and crack problems. *Applicable Anal.* 1984; 18(3):183–219.
- Teltscher M, Maischak M and Stephan EP. A Residual Error Estimator for an Electromagnetic FEM-BEM Coupling Problem in IR³. 2003, submitted. <http://ftp.ifam.uni-hannover.de/pub/preprints/ifam52.ps.Z>.
- Tran T and Stephan EP. Additive Schwarz methods for the h-version boundary element method. *Applicable Anal.* 1996; 60(1–2):63–84.
- Tran T and Stephan EP. Additive Schwarz algorithms for the p version of the Galerkin boundary element method. *Numer. Math.* 2000; 85(3):433–468.
- Wathen AJ, Fischer B and Silvester DJ. The convergence rate of the minimum residual method for the Stokes problem. *Numer. Math.* 1995; 71(1):121–134.
- Wathen AJ and Silvester DJ. Fast iterative solution of stabilised Stokes systems part I: using simple diagonal preconditioners. *SIAM J. Numer. Anal.* 1993; 30(3):630–649.
- Wathen AJ and Stephan EP. *Convergence of Preconditioned Minimum Residual Iteration for Coupled Finite Element/Boundary Element Computations*. Mathematics Research Report AM-94-03, University of Bristol: Bristol, 1994.
- Wendland WL. On asymptotic error estimates for the combined boundary and finite element method. In *Innovative Numerical Methods in Engineering*, Shaw RP, Periaux J, Chaudouet A, Wu J, Marino C and Brebbia CA (eds). Springer-Verlag: Berlin, 1986; 55–69.
- Wendland WL. On asymptotic error estimates for combined FEM and BEM. In *Finite Element and Boundary Element Techniques from Mathematical and Engineering Point of View, CISM Courses, 301*, Stein E and Wendland WL (eds). Springer-Verlag, New York, 1988; 273–333.
- Wendland WL. On the coupling of finite elements and boundary elements. In *Discretization Methods in Structural Mechanics, IUTAM/IACM Symposium 1989*, Kuhn G and Mang H (eds). Springer-Verlag: Berlin, 1990; 405–414.
- Yserentant H. On the multi-level splitting of finite element spaces. *Numer. Math.* 1986; 49(4):379–412.
- Zeidler E. *Nonlinear Functional Analysis and its Applications IV*. Springer: New York, 1988.
- Zienkiewicz OC, Kelly DW and Bettess P. Marriage of the best of both worlds (finite elements and boundary integrals). In *Energy Methods in Finite Element Analysis*, Glowinski R, Rodin EY and Zienkiewicz OZ (eds). John Wiley: Chichester, 1979; 81–107.

Chapter 14

Arbitrary Lagrangian–Eulerian Methods

Jean Donea¹, Antonio Huerta², J.-Ph. Ponthot¹ and A. Rodríguez-Ferran²

¹ Université de Liège, Liège, Belgium

² Universitat Politècnica de Catalunya, Barcelona, Spain

| | |
|--|-----|
| 1 Introduction | 413 |
| 2 Descriptions of Motion | 415 |
| 3 The Fundamental ALE Equation | 417 |
| 4 ALE Form of Conservation Equations | 419 |
| 5 Mesh-update Procedures | 420 |
| 6 ALE Methods in Fluid Dynamics | 422 |
| 7 ALE Methods in Nonlinear Solid Mechanics | 426 |
| References | 433 |

1 INTRODUCTION

The numerical simulation of multidimensional problems in fluid dynamics and nonlinear solid mechanics often requires coping with strong distortions of the continuum under consideration while allowing for a clear delineation of free surfaces and fluid–fluid, solid–solid, or fluid–structure interfaces. A fundamentally important consideration when developing a computer code for simulating problems in this class is the choice of an appropriate kinematical description of the continuum. In fact, such a choice determines the relationship between the deforming continuum and the finite grid or mesh of computing zones, and thus conditions the ability of the numerical method to deal with large distortions and provide an accurate resolution of material interfaces and mobile boundaries.

Encyclopedia of Computational Mechanics, Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes. Volume 1: *Fundamentals*. © 2004 John Wiley & Sons, Ltd. ISBN: 0-470-84699-2.

The algorithms of continuum mechanics usually make use of two classical descriptions of motion: the Lagrangian description and the Eulerian description; see, for instance, Malvern (1969). The arbitrary Lagrangian–Eulerian (ALE, in short) description, which is the subject of the present chapter, was developed in an attempt to combine the advantages of the above classical kinematical descriptions, while minimizing their respective drawbacks as far as possible.

Lagrangian algorithms, in which each individual node of the computational mesh follows the associated material particle during motion (see Figure 1), are mainly used in structural mechanics. The Lagrangian description allows an easy tracking of free surfaces and interfaces between different materials. It also facilitates the treatment of materials with history-dependent constitutive relations. Its weakness is its inability to follow large distortions of the computational domain without recourse to frequent remeshing operations.

Eulerian algorithms are widely used in fluid dynamics. Here, as shown in Figure 1, the computational mesh is fixed and the continuum moves with respect to the grid. In the Eulerian description, large distortions in the continuum motion can be handled with relative ease, but generally at the expense of precise interface definition and the resolution of flow details.

Because of the shortcomings of purely Lagrangian and purely Eulerian descriptions, a technique has been developed that succeeds, to a certain extent, in combining the best features of both the Lagrangian and the Eulerian approaches. Such a technique is known as the *arbitrary Lagrangian–Eulerian (ALE) description*. In the ALE description, the nodes of the computational mesh may be moved with the continuum in normal Lagrangian fashion, or be held fixed in Eulerian manner, or, as suggested in

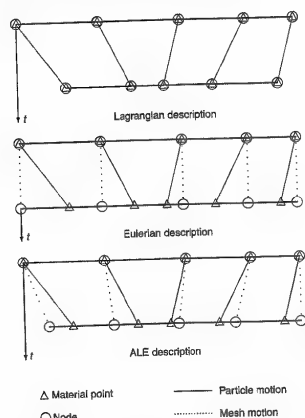


Figure 1. One-dimensional example of Lagrangian, Eulerian and ALE mesh and particle motion.

Figure 1, be moved in some arbitrarily specified way to give a continuous rezoning capability. Because of this freedom in moving the computational mesh offered by the ALE description, greater distortions of the continuum can be handled than would be allowed by a purely Lagrangian method, with more resolution than that afforded by a purely Eulerian approach. The simple example in Figure 2 illustrates the ability of the ALE description to accommodate significant distortions of the computational mesh, while preserving the clear delineation of interfaces typical of a purely Lagrangian approach. A coarse finite element mesh is used to model the detonation of an explosive charge in an extremely strong cylindrical vessel partially filled with water. A comparison is made of the mesh configurations at time $t = 1.0$ ms obtained respectively, with the ALE description (with automatic continuous rezoning) and with a purely Lagrangian mesh description. As further evidenced by the details of the charge-water interface, the Lagrangian approach suffers from a severe degradation of the computational mesh, in contrast with the ability of the ALE approach to maintain quite a regular mesh configuration of the charge-water interface.

The aim of the present chapter is to provide an in-depth survey of ALE methods, including both conceptual

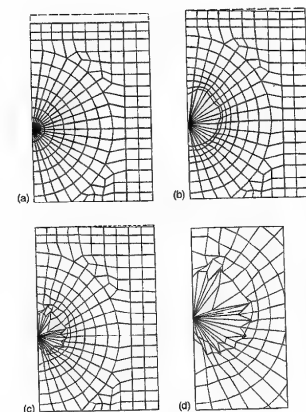


Figure 2. Lagrangian versus ALE descriptions: (a) initial FE mesh; (b) ALE mesh at $t = 1$ ms; (c) Lagrangian mesh at $t = 1$ ms; (d) details of interface in Lagrangian description.

aspects and numerical implementation details in view of the applications in large deformation material response, fluid dynamics, nonlinear solid mechanics, and coupled fluid-structure problems. The chapter is organized as follows. The next section introduces the ALE kinematical description as a generalization of the classical Lagrangian and Eulerian descriptions of motion. Such generalization rests upon the introduction of a so-called *referential domain* and on the mapping between the referential domain and the classical, material, and spatial domains. Then, the fundamental ALE equation is introduced, which provides a relationship between material time derivative and referential time derivative. On this basis, the ALE form of the basic conservation equations for mass, momentum, and energy is established. Computational aspects of the ALE algorithms are then addressed. This includes mesh-update procedures in finite element analysis, the combination of ALE and mesh-refinement procedures, as well as the use of ALE in connection with mesh-free methods. The chapter closes with a discussion of problems commonly encountered in the computer implementation of ALE algorithms in fluid dynamics, solid mechanics, and coupled problems describing fluid-structure interaction.

2 DESCRIPTIONS OF MOTION

Since the ALE description of motion is a generalization of the Lagrangian and Eulerian descriptions, we start with a brief reminder of these classical descriptions of motion. We closely follow the presentation by Donea and Huerta (2003).

2.1 Lagrangian and Eulerian viewpoints

Two domains are commonly used in continuum mechanics: the material domain $R_X \subset \mathbb{R}^{n_d}$, with n_d spatial dimensions, made up of material particles X , and the spatial domain R_x , consisting of spatial points x .

The Lagrangian viewpoint consists of following the material particles of the continuum in their motion. To this end, one introduces, as suggested in Figure 3, a computational grid, which follows the continuum in its motion, the grid nodes being permanently connected to the same material points. The material coordinates, X , allow us to identify the reference configuration, R_X . The motion of the material points relates the material coordinates, X , to the spatial ones, x . It is defined by an application φ such that

$$\varphi: R_X \times [t_0, t_{\text{final}}] \rightarrow R_x \times [t_0, t_{\text{final}}] \\ (X, t) \mapsto \varphi(X, t) = (x, t) \quad (1)$$

which allows us to link X and x in time by the law of motion, namely

$$x = x(X, t), \quad t = t \quad (2)$$

which explicitly states the particular nature of φ : first, the spatial coordinates x depend both on the material particle, X , and time t , and, second, physical time is measured by the same variable t in both material and spatial domains. For every fixed instant t , the mapping φ defines a configuration in the spatial domain. It is convenient to employ a matrix representation for the gradient of φ ,

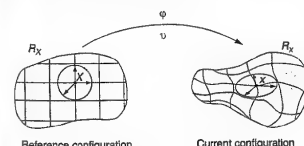


Figure 3. Lagrangian description of motion.

$$\frac{\partial \varphi}{\partial (X, t)} = \begin{pmatrix} \frac{\partial x}{\partial X} & v \\ 0^T & 1 \end{pmatrix} \quad (3)$$

where 0^T is a null row-vector and the material velocity v is

$$v(X, t) = \frac{\partial x}{\partial t} \Big|_X \quad (4)$$

with $|_X$ meaning "holding the material coordinate X fixed".

Obviously, the one-to-one mapping φ must verify $\det(\partial x / \partial X) > 0$ (nonzero to impose a one-to-one correspondence and positive to avoid orientation change of the reference axes) at each point X and instant $t > t_0$. This allows us to keep track of the history of motion and, by the inverse transformation $(X, t) = \varphi^{-1}(x, t)$, to identify, at any instant, the initial position of the material particle occupying position x at time t .

Since the material points coincide with the same grid points during the whole motion, there are no convective effects in Lagrangian calculations: the material derivative reduces to a simple time derivative. The fact that each finite element of a Lagrangian mesh always contains the same material particles represents a significant advantage from the computational viewpoint, especially in problems involving materials with history-dependent behavior. This aspect is discussed in detail by Bonet and Wood (1997). However, when large material deformations do occur, for instance vortices in fluids, Lagrangian algorithms undergo a loss in accuracy, and may even be unable to conclude a calculation, due to excessive distortions of the computational mesh linked to the material.

The difficulties caused by an excessive distortion of the finite element grid are overcome in the Eulerian formulation. The basic idea in the Eulerian formulation, which is very popular in fluid mechanics, consists in examining, as time evolves, the physical quantities associated with the fluid particles passing through a fixed region of space. In an Eulerian description, the finite element mesh is thus fixed and the continuum moves and deforms with respect to the computational grid. The conservation equations are formulated in terms of the spatial coordinates x and the time t . Therefore, the Eulerian description of motion only involves variables and functions having an instantaneous significance in a fixed region of space. The material velocity v at a given mesh node corresponds to the velocity of the material point coincident at the considered time t with the considered node. The velocity v is consequently expressed with respect to the fixed-element mesh without any reference to the initial configuration of the continuum and the material coordinates X : $v = v(x, t)$.

Since the Eulerian formulation dissociates the mesh nodes from the material particles, convective effects appear because of the relative motion between the deforming material and the computational grid. Eulerian algorithms present numerical difficulties due to the nonsymmetric character of convection operators, but permit an easy treatment of complex material motion. By contrast with the Lagrangian description, serious difficulties are now found in following deforming material interfaces and mobile boundaries.

2.2 ALE kinematical description

The above reminder of the classical Lagrangian and Eulerian descriptions has highlighted the advantages and drawbacks of each individual formulation. It has also shown the potential interest in a generalized description capable of combining at best the interesting aspects of the classical mesh descriptions while minimizing their drawbacks as far as possible. Such a generalized description is termed *arbitrary Lagrangian-Eulerian* (ALE) description. ALE methods were first proposed in the finite difference and finite volume context. Original developments were made, among others, by Noh (1964), Franck and Lazarus (1964), Trulio (1966), and Hirt *et al.* (1974); this last contribution has been reprinted in 1997. The method was subsequently adopted in the finite element context and early applications are to be found in the work of Donea *et al.* (1977), Belytschko *et al.* (1978), Belytschko and Kennedy (1978), and Hughes *et al.* (1981).

In the ALE description of motion, neither the material configuration R_X nor the spatial configuration R_x is taken as the reference. Thus, a third domain is needed: the referential configuration R_χ where reference coordinates χ are introduced to identify the grid points. Figure 4 shows

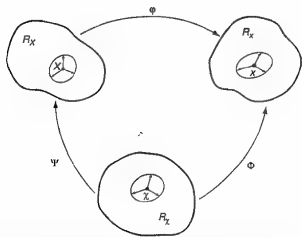


Figure 4. The motion of the ALE computational mesh is independent of the material motion.

these domains and the one-to-one transformations relating the configurations. The referential domain R_χ is mapped into the material and spatial domains by Ψ and Φ respectively. The particle motion φ may then be expressed as $\varphi = \Phi \circ \Psi^{-1}$, clearly showing that, of course, the three mappings Ψ , Φ , and φ are not independent.

The mapping of Φ from the referential domain to the spatial domain, which can be understood as the motion of the grid points in the spatial domain, is represented by

$$\Phi: R_\chi \times [t_0, t_{\text{final}}] \longrightarrow R_x \times [t_0, t_{\text{final}}] \quad (5)$$

$$(\chi, t) \longmapsto \Phi(\chi, t) = (x, t)$$

and its gradient is

$$\frac{\partial \Phi}{\partial (\chi, t)} = \begin{pmatrix} \frac{\partial x}{\partial \chi} & \hat{v} \\ 0^T & 1 \end{pmatrix} \quad (6)$$

where now, the mesh velocity

$$\hat{v}(\chi, t) = \frac{\partial x}{\partial t} \Big|_\chi \quad (7)$$

is involved. Note that both the material and the mesh move with respect to the laboratory. Thus, the corresponding material and mesh velocities have been defined by deriving the equations of material motion and mesh motion respectively with respect to time (see equations 4 and 7).

Finally, regarding Ψ , it is convenient to represent directly its inverse Ψ^{-1} ,

$$\Psi^{-1}: R_X \times [t_0, t_{\text{final}}] \longrightarrow R_\chi \times [t_0, t_{\text{final}}] \quad (8)$$

$$(X, t) \longmapsto \Psi^{-1}(X, t) = (\chi, t)$$

and its gradient is

$$\frac{\partial \Psi^{-1}}{\partial (X, t)} = \begin{pmatrix} \frac{\partial \chi}{\partial X} & w \\ 0^T & 1 \end{pmatrix} \quad (9)$$

where the velocity w is defined as

$$w = \frac{\partial \chi}{\partial t} \Big|_X \quad (10)$$

and can be interpreted as the particle velocity in the referential domain, since it measures the time variation of the referential coordinate χ holding the material particle X fixed. The relation between velocities \hat{v} , \hat{w} , and w can be obtained by differentiating $\varphi = \Phi \circ \Psi^{-1}$,

$$\frac{\partial \varphi}{\partial (X, t)}(X, t) = \frac{\partial \Phi}{\partial (X, t)}(\Psi^{-1}(X, t)) \frac{\partial \Psi^{-1}}{\partial (X, t)}(X, t) \quad (11)$$

$$= \frac{\partial \Phi}{\partial (X, t)}(\chi, t) \frac{\partial \Psi^{-1}}{\partial (X, t)}(X, t)$$

or, in matrix format:

$$\begin{pmatrix} \frac{\partial x}{\partial X} & v \\ 0^T & 1 \end{pmatrix} \begin{pmatrix} \frac{\partial x}{\partial \chi} & \hat{v} \\ 0^T & 1 \end{pmatrix} \begin{pmatrix} \frac{\partial \chi}{\partial X} & w \\ 0^T & 1 \end{pmatrix} \quad (12)$$

which yields, after block multiplication,

$$v = \hat{v} + \frac{\partial x}{\partial \chi} \cdot w \quad (13)$$

This equation may be rewritten as

$$c := v - \hat{v} = \frac{\partial x}{\partial \chi} \cdot w \quad (14)$$

thus defining the convective velocity c , that is, the relative velocity between the material and the mesh.

The convective velocity c (see equation 14), should not be confused with w (see equation 10). As stated before, w is the particle velocity as seen from the referential domain R_χ , whereas c is the particle velocity relative to the mesh as seen from the spatial domain R_x (both v and \hat{v} are variations of coordinate x). In fact, equation (14) implies that $c = w$ if and only if $\partial x / \partial \chi = I$ (where I is the identity tensor), that is, when the mesh motion is purely translational, without rotations or deformations of any kind.

After the fundamentals on ALE kinematics have been presented, it should be remarked that both Lagrangian or Eulerian formulations may be obtained as particular cases. With the choice $\Psi = I$, equation (3) reduces to $X = \chi$ and a Lagrangian description results: the material and mesh velocities, equations (4) and (7), coincide, and the convective velocity c (see equation 14), is null (there are no convective terms in the conservation laws). If, on the other hand, $\Phi = I$, equation (2) simplifies into $x = \chi$, thus implying a Eulerian description: a null mesh velocity is obtained from equation (7) and the convective velocity c is simply identical to the material velocity v .

In the ALE formulation, the freedom of moving the mesh is very attractive. It helps to combine the respective advantages of the Lagrangian and Eulerian formulations. This could, however, be overshadowed by the burden of specifying grid velocities well suited to the particular problem under consideration. As a consequence, the practical implementation of the ALE description requires that an automatic mesh-displacement prescription algorithm be supplied.

3 THE FUNDAMENTAL ALE EQUATION

In order to express the conservation laws for mass, momentum, and energy in an ALE framework, a relation between material (or total) time derivative, which is inherent in conservation laws, and referential time derivative is needed.

3.1 Material, spatial, and referential time derivatives

In order to relate the time derivative in the material, spatial, and referential domains, let a scalar physical quantity be described by $f(x, t)$, $f^*(\chi, t)$, and $f^{**}(X, t)$ in the spatial, referential, and material domains respectively. Stars are employed to emphasize that the functional forms are, in general, different.

Since the particle motion φ is a mapping, the spatial description $f(x, t)$, and the material description $f^{**}(X, t)$ of the physical quantity can be related as

$$f^{**}(X, t) = f(\varphi(X, t), t) \quad \text{or} \quad f^{**} = f \circ \varphi \quad (15)$$

The gradient of this expression can be easily computed as

$$\frac{\partial f^{**}}{\partial (X, t)}(X, t) = \frac{\partial f}{\partial (x, t)}(x, t) \frac{\partial \varphi}{\partial (X, t)}(X, t) \quad (16)$$

which is amenable to the matrix form

$$\begin{pmatrix} \frac{\partial f^{**}}{\partial X} & \frac{\partial f^{**}}{\partial t} \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial t} \end{pmatrix} \begin{pmatrix} \frac{\partial x}{\partial X} & v \\ 0^T & 1 \end{pmatrix} \quad (17)$$

which renders, after block multiplication, a first expression, which is obvious, that is, $(\partial f^{**} / \partial X) = (\partial f / \partial x)(\partial x / \partial X)$; however, the second one is more interesting:

$$\frac{\partial f^{**}}{\partial t} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} \cdot v \quad (18)$$

Note that this is the well-known equation that relates the material and the spatial time derivatives. Dropping the stars to ease the notation, this relation is finally cast as

$$\frac{\partial f}{\partial t} \Big|_X = \frac{\partial f}{\partial t} \Big|_x + v \cdot \nabla f \quad \text{or} \quad \frac{df}{dt} = \frac{\partial f}{\partial t} + v \cdot \nabla f \quad (19)$$

which can be interpreted in the usual way: the variation of a physical quantity for a given particle X is the local variation plus a convective term taking into account the relative motion between the material and spatial (laboratory) systems. Moreover, in order not to overload the rest of the text with notation, except for the specific sections, the material time derivative is denoted as

$$\frac{d}{dt} := \frac{\partial}{\partial t} \Big|_X \quad (20)$$

and the spatial time derivative as

$$\frac{\partial}{\partial t} := \frac{\partial}{\partial t} \Big|_x \quad (21)$$

The relation between material and spatial time derivatives is now extended to include the referential time derivative.

With the help of mapping Ψ , the transformation from the referential description $f^*(X, t)$ of the scalar physical quantity to the material description $f^{**}(X, t)$ can be written as

$$f^{**} = f^* \circ \Psi^{-1} \quad (22)$$

and its gradient can be easily computed as

$$\frac{\partial f^{**}}{\partial X}(X, t) = \frac{\partial f^*}{\partial X}(X, t) \frac{\partial \Psi^{-1}}{\partial X}(X, t) \quad (23)$$

or, in matrix form

$$\left(\frac{\partial f^{**}}{\partial X} \quad \frac{\partial f^{**}}{\partial t} \right) = \left(\frac{\partial f^*}{\partial X} \quad \frac{\partial f^*}{\partial t} \right) \begin{pmatrix} \frac{\partial X}{\partial X} & w \\ 0 & 1 \end{pmatrix} \quad (24)$$

which renders, after block multiplication,

$$\frac{\partial f^{**}}{\partial t} = \frac{\partial f^*}{\partial t} + \frac{\partial f^*}{\partial X} \cdot w \quad (25)$$

Note that this equation relates the material and the referential time derivatives. However, it also requires the evaluation of the gradient of the considered quantity in the referential domain. This can be done, but in computational mechanics it is usually easier to work in the spatial (or material) domain. Moreover, in fluids, constitutive relations are naturally expressed in the spatial configuration and the Cauchy stress tensor, which will be introduced next, is the natural measure for stresses. Thus, using the definition of w given in equation (14), the previous equation may be rearranged into

$$\frac{\partial f^{**}}{\partial t} = \frac{\partial f^*}{\partial t} + \frac{\partial f^*}{\partial X} \cdot c \quad (26)$$

The fundamental ALE relation between material time derivatives, referential time derivatives, and spatial gradient is finally cast as (stars dropped)

$$\frac{\partial f}{\partial t} \Big|_X = \frac{\partial f}{\partial t} \Big|_x + \frac{\partial f}{\partial X} \cdot c = \frac{\partial f}{\partial t} \Big|_x + c \cdot \nabla f \quad (27)$$

and shows that the time derivative of the physical quantity f for a given particle X , that is, its material derivative, is its local derivative (with the reference coordinate X held fixed) plus a convective term taking into account the relative velocity c between the material and the reference system. This equation is equivalent to equation (19) but in the ALE formulation, that is, when (X, t) is the reference.

3.2 Time derivative of integrals over moving volumes

To establish the integral form of the basic conservation laws for mass, momentum, and energy, we also need to consider the rate of change of integrals of scalar and vector functions over a moving volume occupied by fluid.

Consider thus a material volume V_t bounded by a smooth closed surface S_t whose points at time t move with the material velocity $v = v(x, t)$ where $x \in S_t$. A material volume is a volume that permanently contains the same particles of the continuum under consideration. The material time derivative of the integral of a scalar function $f(x, t)$ (note that f is defined in the spatial domain) over the time-varying material volume V_t is given by the following well-known expression, often referred to as Reynolds transport theorem (see, for instance, Belytschko *et al.*, 2000 for a detailed proof):

$$\frac{d}{dt} \int_{V_t} f(x, t) dV = \int_{V_t} \frac{\partial f(x, t)}{\partial t} dV + \int_{S_t=S_t} f(x, t) v \cdot n dS \quad (28)$$

which holds for smooth functions $f(x, t)$. The volume integral in the right-hand side is defined over a control volume V_c (fixed in space), which coincides with the moving material volume V_t at the considered instant, t , in time. Similarly, the fixed control surface S_c coincides at time t with the closed surface S_t bounding the material volume V_t . In the surface integral, n denotes the unit outward normal to the surface S_t at time t , and v is the material velocity of points of the boundary S_t . The first term in the right-hand side of expression (28) is the *local time derivative* of the volume integral. The boundary integral represents the flux of the scalar quantity f across the fixed boundary of the control volume $V_c \equiv V_t$.

Noting that

$$\int_{S_t} f(x, t) v \cdot n dS = \int_{V_c} \nabla \cdot (fv) dV \quad (29)$$

one obtains the alternative form of Reynolds transport theorem:

$$\frac{d}{dt} \int_{V_t} f(x, t) dV = \int_{V_t} \left(\frac{\partial f(x, t)}{\partial t} + \nabla \cdot (fv) \right) dV \quad (30)$$

Similar forms hold for the material derivative of the volume integral of a vector quantity. Analogous formulae can be developed in the ALE context, that is, with a referential time derivative. In this case, however, the characterizing

velocity is no longer the material velocity v , but the grid velocity \hat{v} .

4 ALE FORM OF CONSERVATION EQUATIONS

To serve as an introduction to the discussion of ALE finite element and finite volume models, we establish in this section the differential and integral forms of the conservation equations for mass, momentum, and energy.

4.1 Differential forms

The ALE differential form of the conservation equations for mass, momentum, and energy are readily obtained from the corresponding well-known Eulerian forms

$$\begin{aligned} \text{Mass:} \quad \frac{d\rho}{dt} + \rho \cdot \nabla v &= -\rho \nabla \cdot v \\ \text{Momentum:} \quad \rho \frac{dv}{dt} &= \rho \left(\frac{\partial v}{\partial t} + (v \cdot \nabla) v \right) = \nabla \cdot \sigma + \rho b \\ \text{Energy:} \quad \rho \frac{dE}{dt} &= \rho \left(\frac{\partial E}{\partial t} + v \cdot \nabla E \right) \\ &= \nabla \cdot (\sigma \cdot v) + v \cdot \rho b \end{aligned} \quad (31)$$

where ρ is the mass density, v is the material velocity vector, σ denotes the Cauchy stress tensor, b is the specific body force vector, and E is the specific total energy. Only mechanical energies are considered in the above form of the energy equation. Note that the stress term in the same equation can be rewritten in the form

$$\begin{aligned} \nabla \cdot (\sigma \cdot v) &= \frac{\partial}{\partial x_i} (\sigma_{ij} v_j) = \frac{\partial \sigma_{ij}}{\partial x_i} v_j + \sigma_{ij} \frac{\partial v_j}{\partial x_i} \\ &= (\nabla \cdot \sigma) \cdot v + \sigma : \nabla v \end{aligned} \quad (32)$$

where ∇v is the spatial velocity gradient.

Also frequently used is the balance equation for the internal energy

$$\rho \frac{de}{dt} = \rho \left(\frac{\partial e}{\partial t} + v \cdot \nabla e \right) = \sigma : \nabla^s v \quad (33)$$

where e is the specific internal energy and $\nabla^s v$ denotes the stretching (or strain rate) tensor, the symmetric part of the velocity gradient ∇v ; that is, $\nabla^s v = (1/2)(\nabla v + \nabla^T v)$.

All one has to do to obtain the ALE form of the above conservation equations is to replace in the various convective terms, the material velocity v with the convective

velocity $c = v - \hat{v}$. The result is

$$\begin{aligned} \text{Mass:} \quad \frac{\partial \rho}{\partial t} \Big|_X + c \cdot \nabla \rho &= -\rho \nabla \cdot v \\ \text{Momentum:} \quad \rho \left(\frac{\partial v}{\partial t} \Big|_X + (c \cdot \nabla) v \right) &= \nabla \cdot \sigma + \rho b \\ \text{Total energy:} \quad \rho \left(\frac{\partial E}{\partial t} \Big|_X + c \cdot \nabla E \right) &= \nabla \cdot (\sigma \cdot v) + v \cdot \rho b \\ \text{Internal energy:} \quad \rho \left(\frac{\partial e}{\partial t} \Big|_X + c \cdot \nabla e \right) &= \sigma : \nabla^s v. \end{aligned} \quad (34)$$

It is important to note that the right-hand side of equation (34) is written in classical Eulerian (spatial) form, while the arbitrary motion of the computational mesh is only reflected in the left-hand side. The origin of equations (34) and their similarity with the Eulerian equations (31) have induced some authors to name this method the *quasi-Eulerian* description; see, for instance, Belytschko *et al.* (1980).

Remark (Material acceleration) Mesh acceleration plays no role in the ALE formulation, so, only the material acceleration a , the material derivative of velocity v , is needed, which is expressed in the Lagrangian, Eulerian, and ALE formulation respectively as

$$a = \frac{\partial v}{\partial t} \Big|_X \quad (35a)$$

$$a = \frac{\partial v}{\partial t} \Big|_x + v \cdot \frac{\partial v}{\partial x} \quad (35b)$$

$$a = \frac{\partial v}{\partial t} \Big|_X + c \cdot \frac{\partial v}{\partial x} \quad (35c)$$

Note that the ALE expression of acceleration (35c) is simply a particularization of the fundamental relation (27), taking the material velocity v as the physical quantity f . The first term in the right-hand side of relationships (35b) and (35c) represents the local acceleration, the second term being the convective acceleration.

4.2 Integral forms

The starting point for deriving the ALE integral form of the conservation equations is Reynolds transport theorem (28) applied to an arbitrary volume V_t whose boundary $S_t = \partial V_t$ moves with the mesh velocity \hat{v} :

$$\begin{aligned} \frac{\partial}{\partial t} \int_{V_t} f(x, t) dV &= \int_{V_t} \frac{\partial f(x, t)}{\partial t} \Big|_x dV \\ &+ \int_{S_t} f(x, t) \hat{v} \cdot n dS \end{aligned} \quad (36)$$

where, in this case, we have explicitly indicated that the time derivative in the first term of the right-hand side is a spatial time derivative, as in expression (28). We then successively replace the scalar $f(x, t)$ by the fluid density ρ , momentum ρv , and specific total energy ρE . Similarly, the spatial time derivative $\partial f / \partial t$ is substituted with expressions (31) for the mass, momentum, and energy equation. The end result is the following set of ALE integral forms:

$$\begin{aligned} \frac{\partial}{\partial t} \int_V \rho dV + \int_S \rho c \cdot n dS &= 0 \\ \frac{\partial}{\partial t} \int_V \rho v dV + \int_S \rho v c \cdot n dS &= \int_V (\nabla \cdot \sigma + \rho b) dV \\ \frac{\partial}{\partial t} \int_V \rho E dV + \int_S \rho E c \cdot n dS \\ &= \int_V (\rho \cdot \rho b + \nabla \cdot (\sigma \cdot v)) dV \end{aligned} \quad (37)$$

Note that the integral forms for the Lagrangian and Eulerian mesh descriptions are contained in the above ALE forms. The Lagrangian description corresponds to selecting $\hat{v} = v$ ($c = 0$), while the Eulerian description corresponds to selecting $\hat{v} = 0$ ($c = v$).

The ALE differential and integral forms of the conservation equations derived in the present section will be used as a basis for the spatial discretization of problems in fluid dynamics and solid mechanics.

5 MESH-UPDATE PROCEDURES

The majority of modern ALE computer codes are based on either finite volume or finite element spatial discretizations, the former being popular in the fluid mechanics area, the latter being generally preferred in solid and structural mechanics. Note, however, that the ALE methodology is also used in connection with so-called mesh-free methods (see, for instance, Ponthot and Belytschko, 1998 for an application of the element-free Galerkin method to dynamic fracture problems). In the remainder of this chapter, reference will mainly be made to spatial discretizations produced by the finite element method.

As already seen, one of the main advantages of the ALE formulation is that it represents a very versatile combination of the classical Lagrangian and Eulerian descriptions. However, the computer implementation of the ALE technique requires the formulation of a mesh-update procedure that assigns mesh-node velocities or displacements at each station (time step, or load step) of a calculation. The mesh-update strategy can, in principle, be chosen by the user.

However, the remesh algorithm strongly influences the success of the ALE technique and may represent a big burden on the user if it is not rendered automatic.

Two basic mesh-update strategies may be identified. On one hand, the geometrical concept of *mesh regularization* can be exploited to keep the computational mesh as regular as possible and to avoid mesh entanglement during the calculation. On the other hand, if the ALE approach is used as a *mesh-adaptation* technique, for instance, to concentrate elements in zones of steep solution gradient, a suitable indication of the error is required as a basic input to the remesh algorithm.

5.1 Mesh regularization

The objective of mesh regularization is of a geometrical nature. It consists in keeping the computational mesh as regular as possible during the whole calculation, thereby avoiding excessive distortions and squeezing of the computing zones and preventing mesh entanglement. Of course, this procedure decreases the numerical errors due to mesh distortion.

Mesh regularization requires that updated nodal coordinates be specified at each station of a calculation, either through step displacements, or from current mesh velocities \hat{v} . Alternatively, when it is preferable to prescribe the relative motion between the mesh and the material particles, the *referential velocity* w is specified. In this case, \hat{v} is deduced from equation (13). Usually, in fluid flows, the mesh velocity is interpolated, and in solid problems, the mesh displacement is directly interpolated.

First of all, these mesh-updating procedures are classified depending on whether the *boundary motion is prescribed a priori* or its motion is unknown.

When the motion of the material surfaces (usually the boundaries) is known *a priori*, the mesh motion is also prescribed *a priori*. This is done by defining an adequate mesh velocity in the domain, usually by simple interpolation. In general, this implies a Lagrangian description at the moving boundaries (the mesh motion coincides with the prescribed boundary motion), while a Eulerian formulation (fixed mesh velocity $\hat{v} = 0$) is employed far away from the moving boundaries. A transition zone is defined in between. The interaction problem between a rigid body and a viscous fluid studied by Huerta and Liu (1988a) falls in this category. Similarly, the crack propagation problems discussed by Koh and Haber (1986) and Koh *et al.* (1988), where the crack path is known *a priori*, also allow the use of this kind of mesh-update procedure. Other examples of prescribed mesh motion in nonlinear solid mechanics can be found in the works by Liu *et al.* (1986), Huélin *et al.* (1990), van

Haaren *et al.* (2000), and Rodríguez-Ferran *et al.* (2002), among others.

In all other cases, at least a part of the boundary is a material surface whose position must be tracked at each time step. Thus, a Lagrangian description is prescribed along this surface (or at least along its normal). In the first applications to fluid dynamics (usually free surface flows), ALE degrees of freedom were simply divided into purely Lagrangian ($\hat{v} = v$) or purely Eulerian ($\hat{v} = 0$). Of course, the distortion was thus concentrated in a layer of elements. This is, for instance, the case for numerical simulations reported by Noh (1964), Franck and Lazarus (1964), Hirt *et al.* (1974), and Pracht (1975). Nodes located on moving boundaries were Lagrangian, while internal nodes were Eulerian. This approach was used later for fluid-structure interaction problems by Liu and Chang (1984) and in solid mechanics by Haber (1984) and Haber and Hariandja (1985). This procedure was generalized by Hughes *et al.* (1981) using the so-called Lagrange-Euler matrix method. The referential velocity, w , is defined relative to the particle velocity, v , and the mesh velocity is determined from equation (13). Huerta and Liu (1988b) improved this method avoiding the need to solve any equation for the mesh velocity inside the domain and ensuring an accurate tracking of the material surfaces by solving $w \cdot n = 0$, where n is the unit outward normal, only along the material surfaces. Once the boundaries are known, mesh displacements or velocities inside the computational domain can in fact be prescribed through potential-type equations or interpolations as is discussed next.

In fluid-structure interaction problems, solid nodes are usually treated as Lagrangian, while fluid nodes are treated as described above (fixed or updated according to some simple interpolation scheme). Interface nodes between the solid and the fluid must generally be treated as described in Section 6.1.2. Occasionally they can be treated as Lagrangian (see, for instance, Belytschko and Kennedy, 1978; Belytschko *et al.*, 1980, 1982; Belytschko and Liu, 1985; Argyris *et al.*, 1985; Huerta and Liu, 1988b).

Once the boundary motion is known, several interpolation techniques are available to determine the mesh rezoning in the interior of the domain.

5.1.1 Transfinite mapping method

This method was originally designed for creating a mesh on a geometric region with specified boundaries; see e.g. Gordon and Hall (1973), Haber and Abel (1982), and Eriksson (1985). The general transfinite method describes an approximate surface or volume at a nondenumerable number of points. It is this property that gives rise to the term *transfinite mapping*. In the 2-D case, the transfinite mapping can be made to exactly model all domain

boundaries, and, thus, no geometric error is introduced by the mapping. It induces a very low-cost procedure, since new nodal coordinates can be obtained explicitly once the boundaries of the computational domain have been discretized. The main disadvantage of this methodology is that it imposes restrictions on the mesh topology, as two opposite curves have to be discretized with the same number of elements. It has been widely used by the ALE community to update nodal coordinates; see e.g. Ponthot and Hogge (1991), Yamada and Kikuchi (1993), Gadala and Wang (1998, 1999), and Gadala *et al.* (2002).

5.1.2 Laplacian smoothing and variational methods

As in mesh generation or smoothing techniques, the rezoning of the mesh nodes consists in solving a Laplace (or Poisson) equation for each component of the node velocity or position, so that on a logically regular region the mesh forms lines of equal potential. This method is also sometimes called *elliptic mesh generation* and was originally proposed by Winslow (1963). This technique has an important drawback: in a nonconvex domain, nodes may run outside it. Techniques to preclude this pitfall either increase the computational cost enormously or introduce new terms in the formulation, which are particular to each geometry. Examples based on this type of mesh-update algorithms are presented, among others, by Benson (1989, 1992a,b), Liu *et al.* (1988, 1991), Ghosh and Kikuchi (1991), Chenot and Bellet (1995), and Löhner and Yang (1996). An equivalent approach based on a mechanical interpretation: (non)linear elasticity problem is used by Schreurs *et al.* (1986), Le Tallec and Martin (1996), Belytschko *et al.* (2000), and Armero and Love (2003), while Cescutti *et al.* (1988) minimize a functional quantifying the mesh distortion.

5.1.3 Mesh-smoothing and simple interpolations

In fact, in ALE, it is possible to use any mesh-smoothing algorithm designed to improve the shape of the elements once the topology is fixed. Simple iterative averaging procedures can be implemented where possible; see, for instance, Donea *et al.* (1982), Batina (1991), Trépanier *et al.* (1993), Ghosh and Raju (1996), and Aymone *et al.* (2001). A more robust algorithm (especially in the neighborhood of boundaries with large curvature) was proposed by Giuliani (1982) on the basis of geometric considerations. The goal of this method is to minimize both the squeeze and distortion of each element in the mesh. Donea (1983) and Huerta and Casadei (1994) show examples using this algorithm; Sarate and Huerta (2001) and Hermansson and Hansbo (2003)

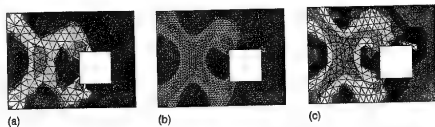


Figure 5. Use of the ALE formulation as an r -adaptive technique. The yield-line pattern is not properly captured with (a) a coarse fixed mesh. Either (b) a fine fixed mesh or (c) a coarse ALE mesh is required. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

made improvements to the original procedure. The main advantage of these mesh-regularization methods is that they are both simple and rather general. They can in fact be applied to unstructured meshes consisting of triangular and quadrilateral elements in 2-D, and to tetrahedral, hexahedral, prism, and pyramidal elements in 3-D.

5.2 Mesh adaptation

When the ALE description is used as an adaptive technique, the objective is to optimize the computational mesh to achieve an improved accuracy, possibly at low computing cost (the total number of elements in a mesh remains unchanged throughout the computation, as well as the element connectivity). Mesh refinement is typically carried out by moving the nodes towards zones of strong solution gradient, such as localization zones in large deformation problems involving softening materials. The ALE algorithm then includes an indicator of the error, and the mesh is modified to obtain an equi-distribution of the error over the entire computational domain. The remesh indicator can, for instance, be made a function of the average or the jump of a certain state variable. Equi-distribution can be carried out using an elliptic or a parabolic differential equation. The ALE technique can nevertheless be coupled with traditional mesh-refinement procedures, such as h -adaptivity, to further enhance accuracy through the selective addition of new degrees of freedom (see Askes and Rodríguez-Ferran, 2001).

Consider, for instance, the use of the ALE formulation for the prediction of yield-line patterns in plates (see Askes *et al.*, 1999). With a coarse fixed mesh, (Figure 5(a)), the spatial discretization is too poor and the yield-line pattern cannot be properly captured. One possible solution is, of course, to use a finer mesh (see Figure 5(b)). Another very attractive possibility from a computational viewpoint is to stay with the coarse mesh and use the ALE formulation to relocate the nodes (see Figure 5(c)). The level of plasticity is used as the remesh indicator. Note that, in this

problem, element distortion is not a concern (contrary to the typical situation illustrated in Figure 2); nodes are relocated to concentrate them along the yield lines.

Studies on the use of ALE as a mesh-adaptation technique in solid mechanics are reported, among others, by Pijaudier-Cabot *et al.* (1995), Huerta *et al.* (1999), Askes and Sluys (2000), Askes and Rodríguez-Ferran (2001), Askes *et al.* (2001), and Rodríguez-Ferran *et al.* (2002).

Mesh adaptation has also found widespread use in fluid dynamics. Often, account must be taken of the directional character of the flow, so anisotropic adaptation procedures are to be preferred. For example, an efficient adaptation method for viscous flows with strong shear layers has to be able to refine directionally to adapt the mesh to the anisotropy of the flow. Anisotropic adaptation criteria again have an error estimate as a basic criterion (see, for instance, Fortin *et al.* (1996), Castro-Díaz *et al.* (1996), Alt-Ali-Yahia *et al.* (2002), Habashi *et al.* (2000), and Müller (2002) for the practical implementation of such procedures).

6 ALE METHODS IN FLUID DYNAMICS

Owing to its superior ability with respect to the Eulerian description to deal with interfaces between materials and mobile boundaries, the ALE description is being widely used for the spatial discretization of problems in fluid and structural dynamics. In particular, the method is frequently employed in the so-called hydrocodes, which are used to simulate the large distortion/deformation response of materials, structures, fluids, and fluid–structure systems. They typically apply to problems in impact and penetration mechanics, fracture mechanics, and detonation/blast analysis. We shall briefly illustrate the specificities of ALE techniques in the modeling of viscous incompressible flows and in the simulation of inviscid, compressible flows, including interaction with deforming structures.

The most obvious influence of an ALE formulation in flow problems is that the convective term must account for the mesh motion. Thus, as already discussed in Section 4.1,

the convective velocity c replaces the material velocity v , which appears in the convective term of Eulerian formulations (see equations 31 and 34). Note that the mesh motion may increase or decrease the convection effects. Obviously, in pure convection (for instance, if a fractional-step algorithm is employed) or when convection is dominant, stabilization techniques must be implemented. The interested reader is urged to consult Chapter 2, Volume 3, for a thorough exposition of stabilization techniques available to remedy the lack of stability of the standard Galerkin formulation in convection-dominated situations, or the textbook by Donea and Huerta (2003).

It is important to note that in standard fluid dynamics, the stress tensor only depends on the pressure and (for viscous flows) on the velocity field at the point and instant under consideration. This is not the case in solid mechanics, as discussed below in Section 7. Thus, stress update is not a major concern in ALE fluid dynamics.

6.1 Boundary conditions

The rest of the discussion of the specificities of the ALE formulation in fluid dynamics concerns boundary conditions. In fact, boundary conditions are related to the problem, not to the description employed. Thus, the same boundary conditions employed in Eulerian or Lagrangian descriptions are implemented in the ALE formulation, that is, along the boundary of the domain, kinematical and dynamical conditions must be defined. Usually, this is formalized as

$$\begin{cases} v = v_D & \text{on } \Gamma_D \\ n \cdot \sigma = t & \text{on } \Gamma_N \end{cases}$$

where v_D and t are the prescribed boundary velocities and tractions respectively; n is the outward unit normal to Γ_N , and Γ_D and Γ_N are the two distinct subsets (Dirichlet and Neumann respectively), which define the piecewise smooth boundary of the computational domain. As usual, stress conditions on the boundaries represent the 'natural boundary conditions', and thus, they are automatically included in the weak form of the momentum conservation equation (see 34).

If part of the boundary is composed of a material surface whose position is unknown, then a mixture of both conditions is required. The ALE formulation allows an accurate treatment of material surfaces. The conditions required on a material surface are: (a) no particles can cross it, and (b) stresses must be continuous across the surface (if a net force is applied to a surface of zero mass, the acceleration is infinite). Two types of material surfaces are discussed here: free surfaces and fluid–structure interfaces, which

may or may not be frictionless (whether or not the fluid is inviscid).

6.1.1 Free surfaces

The unknown position of free surfaces can be computed using two different approaches. First, for the simple case of a single-valued function $z = z(x, y, t)$, a hyperbolic equation must be solved,

$$\frac{\partial z}{\partial t} + (v \cdot \nabla)z = 0$$

This is the kinematic equation of the surface and has been used, for instance, by Ramaswamy and Kawahara (1987), Huerta and Liu (1988b, 1990), and Souli and Zolesio (2001). Second, a more general approach can be obtained by simply imposing the obvious condition that no particle can cross the free surface (because it is a material surface). This can be imposed in a straightforward manner by using a Lagrangian description (i.e. $w = 0$ or $v = \hat{v}$) along this surface. However, this condition may be relaxed by imposing only the necessary condition: w equal to zero along the normal to the boundary (i.e. $n \cdot w = 0$, where n is the outward unit normal to the fluid domain, or $n \cdot v = n \cdot \hat{v}$). The mesh position, normal to the free surface, is determined from the normal component of the particle velocity and remeshing can be performed along the tangent; see, for instance Huerta and Liu (1989) or Braess and Wriggers (2000). In any case, these two alternatives correspond to the kinematical condition; the dynamic condition expresses the stress-free situation, $n \cdot \sigma = 0$, and since it is a homogeneous natural boundary condition, as mentioned earlier, it is directly taken into account by the weak formulation.

6.1.2 Fluid–structure interaction

Along solid-wall boundaries, the particle velocity is coupled to the rigid or flexible structure. The enforcement of the kinematic requirement that no particles can cross the interface is similar to the free-surface case. Thus, conditions $n \cdot w = 0$ or $n \cdot v = n \cdot \hat{v}$ are also used. However, due to the coupling between fluid and structure, extra conditions are needed to ensure that the fluid and structural domains will not detach or overlap during the motion. These coupling conditions depend on the fluid.

For an inviscid fluid (no shear effects), only normal components are coupled because an inviscid fluid is free to slip along the structural interface; that is,

$$\begin{cases} n \cdot u = n \cdot u_s & \text{continuity of normal displacements} \\ n \cdot v = n \cdot v_s & \text{continuity of normal velocities} \end{cases}$$

where the displacement/velocity of the fluid (u/v) along the normal to the interface must be equal to the displacement/velocity of the structure (u_s/v_s) along the same direction. Both equations are equivalent and one or the other is used, depending on the formulation employed (displacements or velocities).

For a viscous fluid, the coupling between fluid and structure requires that velocities (or displacements) coincide along the interface; that is,

$$\begin{cases} u = u_s & \text{continuity of displacements} \\ v = v_s & \text{continuity of velocities} \end{cases}$$

In practice, two nodes are placed at each point of the interface: one fluid node and one structural node. Since the fluid is treated in the ALE formulation, the movement of the fluid mesh may be chosen completely independent of the movement of the fluid itself. In particular, we may constrain the fluid nodes to remain contiguous to the structural nodes, so that all nodes on the sliding interface remain permanently aligned. This is achieved by prescribing the grid velocity \hat{v} of the fluid nodes at the interface to be equal to the material velocity v_s of the adjacent structural nodes. The permanent alignment of nodes at ALE interfaces greatly facilitates the flow of information between the fluid and structural domains and permits fluid-structure coupling to be effected in the simplest and the most elegant manner; that is, the imposition of the previous kinematic conditions is simple because of the node alignment.

The dynamic condition is automatically verified along fixed rigid boundaries, but it presents the classical difficulties in fluid-structure interaction problems when compatibility at nodal level in velocities and stresses is required (both for flexible or rigid structures whose motion is coupled to the fluid flow). This condition requires that the stresses in the fluid be equal to the stresses in the structure. When the behavior of the fluid is governed by the linear Stokes law ($\sigma = -p\mathbf{I} + 2\nu\nabla^s v$) or for inviscid fluids this condition is

$$-pn + 2\nu(n \cdot \nabla^s)v = n \cdot \sigma_s \quad \text{or} \quad -pn = n \cdot \sigma_s$$

respectively, where σ_s is the stress tensor acting on the structure. In the finite element representation, the continuous interface is replaced with a discrete approximation and instead of a distributed interaction pressure, consideration is given to its resultant at each interface node.

There is a large amount of literature on ALE fluid-structure interaction, both for flexible structures and for rigid solids; see, among others, Liu and Chang (1985), Liu and Gvildys (1986), Nomura and Hughes (1992), Le Tallec and Mourou (2001), Casadei *et al.* (2001), Sarrate *et al.* (2001), and Zhang and Hisada (2001).

Remark (Fluid-rigid-body interaction) In some circumstances, especially when the structure is embedded in a fluid and its deformations are small compared with the displacements and rotations of its center of gravity, it is justified to idealize the structure as a rigid body resting on a system consisting of springs and dashpots. Typical situations in which such an idealization is legitimate include the simulation of wind-induced vibrations in high-rise buildings or large bridge girders, the cyclic response of offshore structures exposed to sea currents, as well as the behavior of structures in aeronautical and naval engineering where structural loading and response are dominated by fluid-induced vibrations. An illustrative example of ALE fluid-rigid-body interaction is shown in Section 6.2.

Remark (Normal to a discrete interface) In practice, especially in complex 3-D configurations, one major difficulty is to determine the normal vector at each node of the fluid-structure interface. Various algorithms have been developed to deal with this issue: Casadei and Halleux (1995) and Casadei and Sala (1999) present detailed solutions. In 2-D, the tangent to the interface at a given node is usually defined as parallel to the line connecting the nodes at the ends of the interface segments meeting at that node.

Remark (Free surface and structure interaction) The above discussion of the coupling problem only applies to those portions of the structure that are always submerged during the calculation. As a matter of fact, there may exist portions of the structure, which only come into contact with the fluid some time after the calculation begins. This is, for instance, the case for structural parts above a fluid-free surface. For such portions of the structural domain, some sort of sliding treatment is necessary, as for Lagrangian methods.

6.1.3 Geometric conservation laws

In a series of papers (see Lesoinne and Farhat, 1996; Koobus and Farhat, 1999; Guillard and Farhat, 2000; and Farhat *et al.*, 2001), Farhat and coworkers have discussed the notion of *geometric conservation laws* for unsteady flow computations on moving and deforming finite element or finite volume grids.

The basic requirement is that any ALE computational method should be able to predict exactly the trivial solution of a uniform flow. The ALE equation of mass balance (37)₁ is usually taken as the starting point to derive the geometric conservation law. Assuming uniform fields of density ρ and material velocity v , it reduces to the *continuous geometric conservation law*

$$\frac{\partial}{\partial t} \int_{\Omega} dV = \int_{\partial\Omega} \hat{v} \cdot n \, dS \quad (38)$$

As remarked by Smith (1999), equation (38) can also be derived from the other two ALE integral conservation laws (37) with appropriate restrictions on the flow fields.

Integrating equation (38) in time from t^n to t^{n+1} renders the *discrete geometric conservation law (DGCL)*

$$|\Omega_e^{n+1}| - |\Omega_e^n| = \int_{t^n}^{t^{n+1}} \left(\int_{\partial\Omega_e} \hat{v} \cdot n \, dS \right) dt \quad (39)$$

which states that the change in volume (or area, in 2-D) of each element from t^n to t^{n+1} must be equal to the volume (or area) swept by the element boundary during the time interval. Assuming that the volumes Ω_e in the left-hand side of equation (39) can be computed exactly, this amounts to requiring the exact computation of the flux in the right-hand side also. This poses some restrictions on the update procedure for the grid position and velocity. For instance, Lesoinne and Farhat (1996) show that, for first-order time-integration schemes, the mesh velocity should be computed as $\hat{v}^{n+1/2} = (x^{n+1} - x^n)/\Delta t$. They also point out that, although this intuitive formula was used by many time-integrators prior to DGCLs, it is violated in some instances, especially in fluid-structure interaction problems where mesh motion is coupled with structural deformation.

The practical significance of DGCLs is a debated issue in the literature. As admitted by Guillard and Farhat (2000), 'there are recurrent assertions in the literature stating that, in practice, enforcing the DGCL when computing on moving meshes is unnecessary'. Later, Farhat *et al.* (2001) and other authors have studied the properties of DGCL-enforcing ALE schemes from a theoretical viewpoint. The link between DGCLs and the stability (and accuracy) of ALE schemes is still a controversial topic of current research.

6.2 Applications in ALE fluid dynamics

The first example consists in the computation of cross-flow and rotational oscillations of a rectangular profile. The flow is modeled by the incompressible Navier-Stokes equations and the rectangle is regarded as rigid. The ALE formulation for fluid-rigid-body interaction proposed by Sarrate *et al.* (2001) is used.

Figure 6 depicts the pressure field at two different instants. The flow goes from left to right. Note the cross-flow translation and the rotation of the rectangle. The ALE kinematical description avoids excessive mesh distortion (see Figure 7). For this problem, a computationally efficient rezoning strategy is obtained by dividing the mesh into three zones: (1) the mesh inside the inner circle is prescribed to

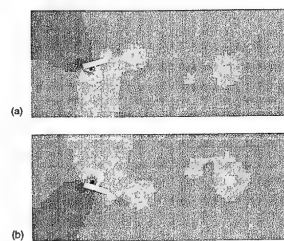


Figure 6. Flow around a rectangle. Pressure fields at two different instants. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ccc>

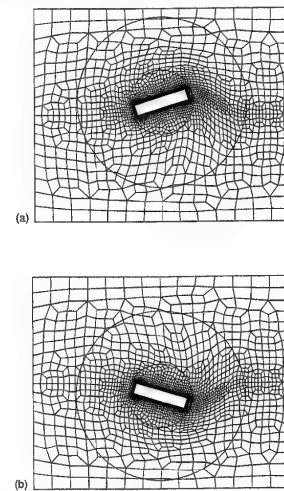


Figure 7. Details of finite element mesh around the rectangle. The ring allows a smooth transition between the rigidly moving mesh around the rectangle and the Eulerian mesh far from it.

move rigidly attached to the rectangle (no mesh distortion and simple treatment of interface conditions); (2) the mesh outside the outer circle is Eulerian (no mesh distortion and no need to select mesh velocity); (3) a smooth transition is prescribed in the ring between the circles (mesh distortion under control).

The second example highlights ALE capabilities for fluid–structure interaction problems. The results shown here, discussed in detail by Casadei and Potapov (2004), have been provided by Casadei and are reproduced here with the authors' kind permission. The example consists in a long 3-D metallic pipe with a square cross section, sealed at both ends, containing a gas at room pressure (see Figure 8). Initially, two 'explosions' take place at the ends of the pipe, simulated by the presence of the same gas, but at a much higher initial pressure.

The gas motion through the pipe is partly affected by internal structures within the pipe (diaphragms #1, #2 and #3) that create a sort of labyrinth. All the pipe walls, and the internal structures, are deformable and are characterized by elastoplastic behavior. The pressures and structural-material properties are so chosen that very large motions and relatively large deformations occur in the structure.

Figure 9 shows the real deformed shapes (not scaled up) of the pipe with superposed fluid-pressure maps. Note the strong wave-propagation effects, the partial wave reflections at obstacles, and the 'ballooning' effect of the thin pipe walls in regions at high pressure. This is a severe test, among other things, for the automatic ALE rezoning algorithms that must keep the fluid mesh reasonably uniform under large motions.

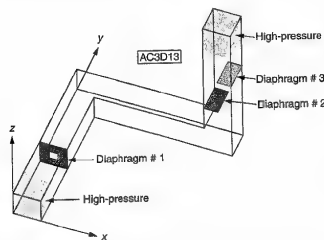


Figure 8. Explosions in a 3-D labyrinth. Problem statement. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

7 ALE METHODS IN NONLINEAR SOLID MECHANICS

Starting in the late 1970s, the ALE formulation has been extended to nonlinear solid and structural mechanics. Particular efforts were made in response to the need to simulate problems describing crack propagation, impact, explosion, vehicle crashes, as well as forming processes of materials. The large distortions/deformations that characterize these problems clearly undermine the utility of the Lagrangian approach traditionally used in problems involving materials with path-dependent constitutive relations. Representative publications on the use of ALE in solid mechanics are, among many others, Liu *et al.* (1986, 1988), Schreurs *et al.* (1986), Benson (1989), Hučink *et al.* (1990), Ghosh and Kikuchi (1991), Baaijens (1993), Huerta and Casadei (1994), Rodríguez-Ferran *et al.* (1998, 2002), Askes *et al.* (1999), and Askes and Sluys (2000).

If mechanical effects are uncoupled from thermal effects, the mass and momentum equations can be solved independently from the energy equation. According to expressions (34), the ALE version of these equations is

$$\frac{\partial \rho}{\partial t} + (c \cdot \nabla) \rho = -\rho \nabla \cdot v \quad (40a)$$

$$\rho a = \rho \frac{\partial v}{\partial t} + \rho (c \cdot \nabla) v = \nabla \cdot \sigma + \rho b \quad (40b)$$

where a is the material acceleration defined in (35a, b and c), σ denotes the Cauchy stress tensor and b represents an applied body force per unit mass.

A standard simplification in nonlinear solid mechanics consists of dropping the mass equation (40a), which is not explicitly accounted for, thus solving only the momentum equation (40b). A common assumption consists of taking the density ρ as constant, so that the mass balance (40a) reduces to

$$\nabla \cdot v = 0 \quad (41)$$

which is the well-known incompressibility condition. This simplified version of the mass balance is also commonly neglected in solid mechanics. This is acceptable because elastic deformations typically induce very small changes in volume, while plastic deformations are volume preserving (isochoric plasticity). This means that changes in density are negligible and that equation (41) automatically holds to sufficient approximation without the need to add it explicitly to the set of governing equations.

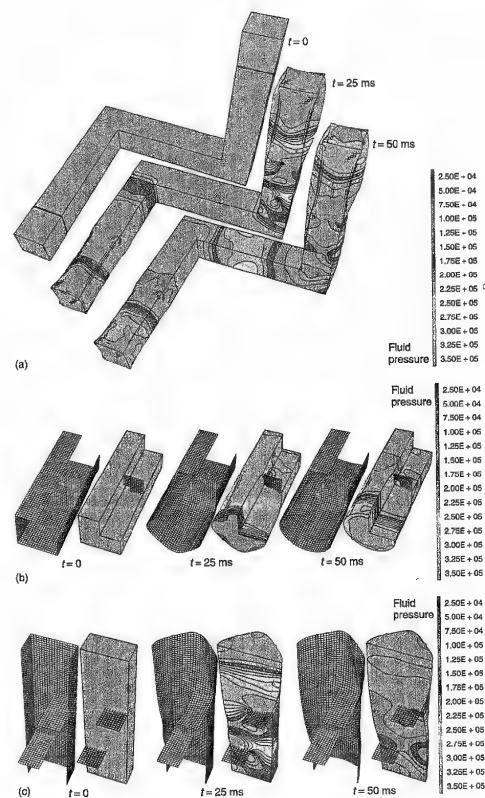


Figure 9. Explosions in a 3-D labyrinth. Deformation in structure and pressure in fluid are properly captured with ALE fluid–structure interaction: (a) whole model; (b) zoom of diaphragm #1; (c) zoom of diaphragms #2 and #3. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

7.1 ALE treatment of steady, quasistatic and dynamic processes

In discussing the ALE form (40b) of the momentum equation, we shall distinguish between steady, quasistatic, and dynamic processes. In fact, the expression for the inertia forces ρa critically depends on the particular type of process under consideration.

A process is called *steady* if the material velocity v in every spatial point x is constant in time. In the Eulerian description (35b), this results in zero local acceleration $\partial v / \partial t|_x$, and only the convective acceleration is present in the momentum balance, which reads

$$\rho a = \rho(v \cdot \nabla)v = \nabla \cdot \sigma + \rho b \quad (42)$$

In the ALE context, it is also possible to assume that a process is steady with respect to a grid point χ and neglect the local acceleration $\partial v / \partial t|_\chi$ in the expression (35c); see, for instance, Ghosh and Kikuchi (1991). The momentum balance then becomes

$$\rho a = \rho(c \cdot \nabla)v = \nabla \cdot \sigma + \rho b \quad (43)$$

However, the physical meaning of a null ALE local acceleration (that is, of an "ALE-steady" process) is not completely clear, due to the arbitrary nature of the mesh velocity and, hence, of the convective velocity c .

A process is termed *quasistatic* if the inertia forces ρa are negligible with respect to the other forces in the momentum balance. In this case, the momentum balance reduces to the static equilibrium equation

$$\nabla \cdot \sigma + \rho b = 0 \quad (44)$$

in which time and material velocity play no role. Since the inertia forces have been neglected, the different descriptions of acceleration in equations (35a, b and c) do not appear in equation (44), which is therefore valid in both Eulerian and ALE formulations. The important conclusion is that there are no convective terms in the ALE momentum balance for quasistatic processes. A process may be modeled as quasistatic if stress variations and/or body forces are much larger than inertia forces. This is a common situation in solid mechanics, encompassing, for instance, various metal-forming processes. As discussed in the next section, convective terms are nevertheless present in the ALE (and Eulerian) constitutive equation for quasistatic processes. They reflect the fact that grid points are occupied by different particles at different times.

Finally, in *transient dynamic processes*, all terms must be retained in expression (35c) for the material acceleration,

and the momentum balance equation is given by expression (40b).

7.2 ALE constitutive equations

Compared to the use of the ALE description in fluid dynamics, the main additional difficulty in nonlinear solid mechanics is the design of an appropriate stress-update procedure to deal with history-dependent constitutive equations. As already mentioned, constitutive equations of ALE nonlinear solid mechanics contain convective terms that account for the relative motion between mesh and material. This is the case for both hypoelastoplastic and hyperelastoplastic models.

7.2.1 Constitutive equations for ALE hypoelastoplasticity

Hypoelastoplastic models are based on an additive decomposition of the stretching tensor $\nabla^s v$ (symmetric part of the velocity gradient) into elastic and plastic parts; see, for instance, Belytschko *et al.* (2000) or Bonet and Wood (1997). They were used in the first ALE formulations for solid mechanics and are still the standard choice. In these models, material behavior is described by a rate-form constitutive equation

$$\sigma^* = f(\sigma, \nabla^s v) \quad (45)$$

relating an objective rate of Cauchy stress σ^* to stress and stretching. The material rate of stress

$$\dot{\sigma} = \frac{\partial \sigma}{\partial t} \Big|_\chi = \frac{\partial \sigma}{\partial t} \Big|_\chi + (c \cdot \nabla)\sigma \quad (46)$$

cannot be employed in relation (45) to measure the stress rate because it is not an objective tensor, so large rigid-body rotations are not properly treated. An objective rate of stress is obtained by adding to $\dot{\sigma}$ some terms that ensure the objectivity of σ^* ; see, for instance, Malvern (1969) or Belytschko *et al.* (2000). Two popular objective rates are the Truesdell rate and the Jaumann rate

$$\sigma^* = \dot{\sigma} - \nabla^w v \cdot \sigma - \sigma \cdot (\nabla^w v)^T \quad (47)$$

where $\nabla^w v = \frac{1}{2}(\nabla v + \nabla^T v)$ is the spin tensor.

Substitution of equation (47) (or similar expressions for other objective stress rates) into equation (45) yields

$$\dot{\sigma} = g(\sigma, \nabla^s v, \dots) \quad (48)$$

where g contains both f and the terms in σ^* , which ensure its objectivity.

In the ALE context, referential time derivatives, not material time derivatives, are employed to represent evolution in time. Combining expression (46) of the material rate of stress and the constitutive relation (48) yields a rate-form constitutive equation for ALE nonlinear solid mechanics

$$\dot{\sigma} = \frac{\partial \sigma}{\partial t} \Big|_\chi + (c \cdot \nabla)\sigma = g \quad (49)$$

where, again, a convective term reflects the motion of material particles relative to the mesh. Note that this relative motion is inherent in ALE kinematics, so the convective term is present in all the situations described in Section 7.1, including quasistatic processes.

Because of this convective effect, the stress update cannot be performed as simply as in the Lagrangian formulation, in which the element Gauss points correspond to the same material particles during the whole calculation. In fact, the accurate treatment of the convective terms in ALE rate-type constitutive equations is a key issue in the accuracy of the formulation, as discussed in Section 7.3.

7.2.2 Constitutive equations for ALE hyperelastoplasticity

Hyperelastoplastic models are based on a multiplicative decomposition of the deformation gradient into elastic and plastic parts, $F = F^e F^p$; see, for instance, Belytschko *et al.* (2000) or Bonet and Wood (1997). They have only very recently been combined with the ALE description (see Rodríguez-Ferran *et al.*, 2002 and Armero and Love, 2003).

The evolution of stresses is not described by means of a rate-form equation, but in closed form as

$$\tau = 2 \frac{dW}{db^e} b^e \quad (50)$$

where $b^e = F^e \cdot (F^e)^T$ is the elastic left Cauchy-Green tensor, W is the free energy function, and $\tau = \det(F) \sigma$ is the Kirchhoff stress tensor.

Plastic flow is described by means of the flow rule

$$\dot{b}^e - \nabla v \cdot b^e - b^e \cdot (\nabla v)^T = -2\dot{\gamma} m(\tau) \cdot b^e \quad (51)$$

The left-hand side of equation (51) is the Lie derivative of b^e with respect to the material velocity v . In the right-hand side, m is the flow direction and $\dot{\gamma}$ is the plastic multiplier.

Using the fundamental ALE relation (27) between material and referential time derivatives, the flow rule (51) can be recast as

$$\frac{\partial b^e}{\partial t} \Big|_\chi + (c \cdot \nabla)b^e = \nabla v \cdot b^e + b^e \cdot (\nabla v)^T - 2\dot{\gamma} m(\tau) \cdot b^e \quad (52)$$

Note that, like in equation (49), a convective term in this constitutive equation reflects the relative motion between mesh and material.

7.3 Stress-update procedures

In the context of hypoelastoplasticity, various strategies have been proposed for coping with the convective terms in equation (49). Following Benson (1992b), they can be classified into *split* and *unsplit* methods.

If an *unsplit* method is employed, the complete rate equation (49) is integrated forward in time, including both the convective term and the material term g . This approach is followed, among others, by Liu *et al.* (1986), who employed an explicit time-stepping algorithm and by Ghosh and Kikuchi (1991), who used an implicit unsplit formulation.

On the other hand, *split*, or fractional-step methods treat the material and convective terms in (49) in two distinct phases: a material (or Lagrangian) phase is followed by a convection (or transport) phase. In exchange for a certain loss in accuracy due to splitting, split methods are simpler and especially suitable in upgrading a Lagrangian code to the ALE description. An implicit split formulation is employed by Huétink *et al.* (1990) to model metal-forming processes. An example of explicit split formulation may be found in Huerta and Casadei (1994), where ALE finite elements are used to model fast-transient phenomena.

The situation is completely analogous for hyperelastoplasticity, and similar comments apply to the split or unsplit treatment of material and convective effects. In fact, if a split approach is chosen (see Rodríguez-Ferran *et al.*, 2002), the only differences with respect to the hypoelastoplastic models are (1) the constitutive model for the Lagrangian phase (hypo/hyper) and (2) the quantities to be transported in the convection phase.

7.3.1 Lagrangian phase

In the Lagrangian phase, convective effects are neglected. The constitutive equations recover their usual expressions (48) and (51) for hypo- and hyper-models respectively. The ALE kinematical description has (momentarily) disappeared from the formulation, so all the concepts, ideas, and algorithms of large strain solid mechanics with a Lagrangian description apply (see Bonet and Wood (1997), Belytschko *et al.* (2000), and Chapter 7, Volume 2).

The issue of objectivity is one of the main differences between hypo- and hypermodels. When devising time-integration algorithms to update stresses from σ^n to σ^{n+1}

in hypoelastoplastic models, a typical requirement is incremental objectivity (that is, the appropriate treatment of rigid-body rotations over the time interval $[t^n, t^{n+1}]$). In hyperelastoplastic models, on the contrary, objectivity is not an issue at all, because there is no rate equation for the stress tensor.

7.3.2 Convection phase

The convective effects neglected before have to be accounted for now. Since material effects have already been treated in the Lagrangian phase, the ALE constitutive equations read simply

$$\frac{\partial \sigma}{\partial t} \Big|_X + (c \cdot \nabla) \sigma = 0; \quad \frac{\partial b^e}{\partial t} \Big|_X + (c \cdot \nabla) b^e = 0; \quad (53)$$

$$\frac{\partial \alpha}{\partial t} \Big|_X + (c \cdot \nabla) \alpha = 0 \quad (53)$$

Equations (53)₁ and (53)₂ correspond to hypo- and hyperelastoplastic models respectively (cf. with equations 49 and 52). In equation (53)₃, valid for both hypo- and hyper-models, α is the set of all the material-dependent variables, i.e. variables associated with the material particle X : internal variables for hardening or softening plasticity, the volume change in nonisochoric plasticity, and so on (see Rodríguez-Ferran *et al.*, 2002).

The three equations in (53) can be written more compactly as

$$\frac{\partial \blacksquare}{\partial t} \Big|_X + (c \cdot \nabla) \blacksquare = 0 \quad (54)$$

where \blacksquare represents the appropriate variable in each case. Note that equation (54) is simply a first-order linear hyperbolic PDE, which governs the transport of field \blacksquare by the velocity field c . However, two important aspects should be considered in the design of numerical algorithms for the solution of this equation:

1. \blacksquare is a tensor (for σ and b^e) or vector-like (for α) field, so equation (54) should be solved for each component \square of \blacksquare :

$$\frac{\partial \square}{\partial t} \Big|_X + c \cdot \nabla \square = 0 \quad (55)$$

Since the number of scalar equations (55) may be relatively large (for instance: eight for a 3-D computation with a plastic model with two internal variables), the need for efficient convection algorithms is a key issue in ALE nonlinear solid mechanics.

2. \square is a Gauss-point-based (i.e. not a nodal-based) quantity, so it is discontinuous across finite element edges. For this reason, its gradient $\nabla \square$ cannot be

reliably computed at the element level. In fact, handling $\nabla \square$ is the main numerical challenge in ALE stress update.

Two different strategies may be used to tackle the difficulties associated with $\nabla \square$. One possible approach is to approximate \square by a continuous field $\bar{\square}$, and replace $\nabla \square$ by $\nabla \bar{\square}$ in equation (55). The smoothed field $\bar{\square}$ can be obtained, for instance, by least-squares approximation (see Huétink *et al.*, 1990).

Another possibility is to retain the discontinuous field \square and devise appropriate algorithms that account for this fact. To achieve this aim, a fruitful observation is noting that, for a piecewise constant field \square , equation (55) is the well-known Riemann problem. Through this connection, the ALE community has exploited the expertise on approximate Riemann solvers of the CFD community (see Le Veque, 1990).

Although \square is, in general, not constant for each element (except for one-point quadratures), it can be approximated by a piecewise constant field in a simple manner. Figure 10 depicts a four-noded quadrilateral with a 2×2 quadrature subdivided into four subelements. If the value of \square for each Gauss point is taken as representative for the whole subelement, then a field constant within each subelement results.

In this context, equation (55) can be solved explicitly by looping all the subelements in the mesh by means of a Godunov-like technique based on Godunov's method for conservation laws (see Rodríguez-Ferran *et al.*, 1998):

$$\square^{n+1} = \square^n - \frac{\Delta t}{V} \sum_{r=1}^{N_r} f_r (\square_r^n - \square^n) [1 - \text{sign}(f_r)] \quad (56)$$

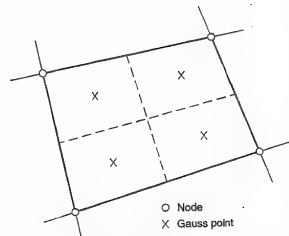


Figure 10. Finite element subdivided into subelements for the Godunov-like stress update.

According to equation (56), the Lagrangian (i.e. after the Lagrangian phase) value \square^n is updated into the final value \square^{n+1} by taking into account the flux of \square across the subelement edges. In the second term on the right-hand side, Δt is the time step, V is the volume (or area, in 2-D) of the subelement, N_r is the number of edges per subelement, \square_r^n is the value of \square in the contiguous subelement across edge r , and f_r is the flux of convective velocity across edge r , $f_r = \int_r (c \cdot n) d\Gamma$. Note that a full-donor (i.e. full upwind) approach is obtained by means of $\text{sign}(f_r)$.

Remark (Split stress update and iterations) In principle, the complete stress update must be performed at each iteration of the nonlinear equilibrium problem. However, a common simplification consists in leaving the convection phase outside the iteration loop (see Baaijens, 1993; Rodríguez-Ferran *et al.*, 1998; and Rodríguez-Ferran *et al.*, 2002); that is, iterations are performed in a purely Lagrangian fashion up to equilibrium and the convection phase is performed after remeshing, just once per time step. Numerical experiments reveal that disruption of equilibrium caused by convection is not severe and can be handled as extra residual forces in the next load step (see references just cited).

Remark (ALE finite strain elasticity) Since hyperelasticity can be seen as a particular case of hyperelastoplasticity, we have chosen here the more general situation for a more useful presentation. In the particular case of large elastic strains (hyperelasticity), where $\mathbf{F} = \mathbf{F}^e$, an obvious option is to particularize the general approach just described by solving only the relevant equation (i.e. equation 52 for b^e ; note that there are no internal variables α in elasticity). In the literature, other approaches exist for this specific case. Yamada and Kikuchi (1993) and Armero and Love (2003) exploit the relationship $\mathbf{F} = \mathbf{F}_\Phi \mathbf{F}_\Psi^{-1}$, where \mathbf{F}_Φ and \mathbf{F}_Ψ respectively (see Figure 4), to obtain an ALE formulation for hyperelasticity with no convective terms. In exchange for the need to handle the convective term in the update of b^e (see equations 52 and 53), the former approach has the advantage that only the quality of mapping Φ needs to be controlled. In the latter approaches, on the contrary, both the quality of Φ and Ψ must be ensured; that is, two meshes (instead of only one) must be controlled.

7.4 Applications in ALE nonlinear solid mechanics

For illustrative purposes, two powder compaction ALE simulations are briefly discussed here. More details can be found in Rodríguez-Ferran *et al.* (2002) and Pérez-Foguet *et al.* (2003).

The first example involves the bottom punch compaction of an axisymmetric flanged component (see Figure 11). If a Lagrangian description is used (see Figure 11(a)), the large upward mass flow leads to severe element distortion in the reentrant corner, which in turn affects the accuracy in the relative density. With an ALE description (see Figure 11(b)), mesh distortion is completely precluded by means of a very simple ALE remeshing strategy: a uniform vertical mesh compression is prescribed in the bottom, narrow part of the piece, and no mesh motion (i.e. Eulerian mesh) in the upper, wide part.

The second example involves the compaction of a multilevel component. Owing to extreme mesh distortion, it is not possible to perform the simulation with a Lagrangian approach. Three ALE simulations are compared in Figure 12, corresponding to top, bottom, and simultaneous top-bottom compaction. Even with an unstructured mesh and a more complex geometry, the ALE description avoids mesh distortion, so the final relative density profiles can be computed (see Figure 13).

7.5 Contact algorithms

Contact treatment, especially when frictional effects are present, is a very important feature of mechanical modeling and certainly remains one of the more challenging problems in computational mechanics. In the case of the classical Lagrangian formulation, much attention has been devoted to contact algorithms and the interested reader is referred to references by Chapter 6, Volume 2, Zhong (1993), Wriggers (2002), and Laursen (2002) in order to get acquainted with the subject. By contrast, modeling of contact in connection with the ALE description has received much less attention. Paradoxically, one of the interesting features of ALE is that, in some situations, the formulation can avoid the burden of implementing cumbersome contact algorithms. The coning problem in Figure 14 is a good illustration of the versatility of ALE algorithms in the treatment of frictionless contact over a known (moving) surface. The material particle M located at the punch corner at time t has been represented. At time $t + \Delta t$, the punch has moved slightly downwards and, due to compression, the material particle M has moved slightly to the right. When a Lagrangian formulation is used in the simulation of such a process, due to the fact that the mesh sticks to the material, a contact algorithm has to be used to obtain a realistic simulation. On the contrary, using an ALE formalism, the implementation of a contact algorithm can be avoided. This is simply because the ALE formulation allows us to prevent the horizontal displacement of the mesh nodes located under the

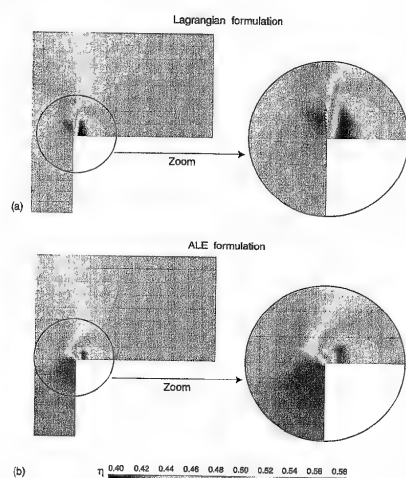


Figure 11. Final relative density after the bottom punch compaction of a flanged component: (a) Lagrangian approach leads to severe mesh distortion; (b) ALE approach avoids distortion. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

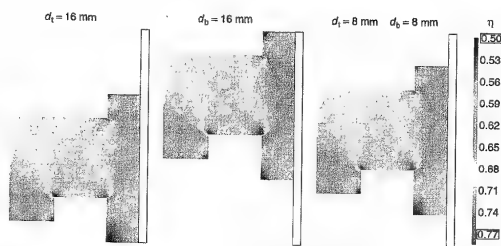


Figure 12. Final relative density of a multilevel component. From left to right: top, bottom and double compaction. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

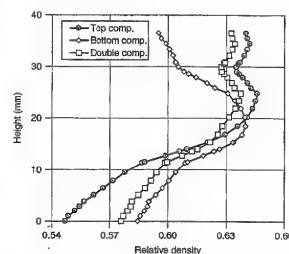


Figure 13. Relative density profiles in a multilevel component along a vertical line for the three-compaction processes.

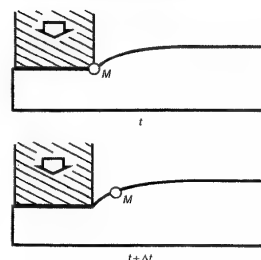


Figure 14. Schematic description of the coining process.

punch, irrespective of the material flow. Numerous illustrations of this particular case can be found, amongst others, in Schreurs *et al.* (1986), Huétink *et al.* (1990), Hogge and Ponthot (1991a), Huerta and Casadei (1994), Rodríguez-Ferran *et al.* (2002), Gadala and Wang (1998), Gadala *et al.* (2002), and Martinet and Chabrand (2000).

However, in more general situations, a contact algorithm cannot be avoided. In such a case, the nodes of contact elements have to be displaced convectively in accordance with the nodes on the surface of the bulk materials and tools. A direct consequence of this displacement is that convective effects have to be taken into account for history-dependent variables. In the simplest penalty case in which the normal pressure is proportional to the penetration, the

normal contact stress depends only on the current geometry and not on the history of penetration. As a consequence, no convection algorithm needs to be activated for that quantity. On the contrary, for a Coulomb-friction model, the shear stresses in the contact elements are incrementally calculated. They therefore depend on the history and hence a convective increment of the shear stress should be evaluated. One simple way to avoid the activation of the convection algorithm for the contact/friction quantities is indeed to keep the boundaries purely Lagrangian. However, in general this will not prevent mesh distortions and nonconvexity in the global mesh.

Various applications of the ALE technology have been developed so far to treat the general case of moving boundaries with Coulomb or Tresca frictional contact. For example, in their pioneering work on ALE contact, Haber and Hariandja (1985) update the mesh, so that nodes and element edges on the surface of contacting bodies coincide exactly at all points along the contact interface in the deformed configuration. In such a case, the matching of node pairs and element edges ensures a precise satisfaction of geometric compatibility and allows a consistent transfer of contact stresses between the two bodies. A similar procedure was established by Ghosh (1992) but, in this case, the numerical model introduces ALE nodal points on one of the contacting (slave) surfaces that are constrained to follow Lagrangian nodes on the other (master) surface. Liu *et al.* (1991) presented an algorithm, mostly dedicated to rolling applications, where the stick nodes are assumed Lagrangian, whereas the slip nodes are equally spaced between the two adjacent stick nodes. More general procedures have been introduced by Huétink *et al.* (1990).

More recently, sophisticated frictional models incorporating lubrication models have been used in ALE formulations. In such a case, the friction between the contacting lubricated bodies is expressed as a function of interface variables (mean lubrication film thickness, sheet and tooling roughness) in addition to more traditional variables (interface pressure, sliding speed, and strain rate). Examples of complex lubrication models integrated into an ALE framework have been presented by Hu and Liu (1992, 1993, 1994), Martinet and Chabrand (2000), and Boman and Ponthot (2002).

REFERENCES

- Ait-Ali-Yahia D., Baruzzi G., Habashi WG., Fortin M., Dompiere J. and Vallet M-G Anisotropic mesh adaptation: towards user-independent, mesh-independent and solver-independent CFD. Part II. Structured grids. *Int. J. Numer. Methods Fluids* 2002; 39(8):657-673.

- Argyris JH, Doltsinis JS, Fisher H and Wüstenberg H. TA PANTIA PEL. *Comput. Methods Appl. Mech. Eng.* 1985; 51(1-3):289-362.
- Armero F and Love E. An arbitrary Lagrangian-Eulerian finite element method for finite strain plasticity. *Int. J. Numer. Methods Eng.* 2003; 57(4):471-508.
- Askes H and Rodríguez-Ferran A. A combined *rh*-adaptive scheme based on domain subdivision. Formulation and linear examples. *Int. J. Numer. Methods Eng.* 2001; 51(3):253-273.
- Askes H, Rodríguez-Ferran A and Huerta A. Adaptive analysis of yield line patterns in plates with the arbitrary Lagrangian-Eulerian method. *Comput. Struct.* 1999; 70(3):257-271.
- Askes H and Sluys LJ. Remeshing strategies for adaptive ALE analysis of strain localisation. *Eur. J. Mech., A-Solids* 2000; 19(3):447-467.
- Askes H, Sluys LJ and de Jong BBC. Remeshing techniques for *r*-adaptive and combined *h/r*-adaptive analysis with application to 2D/3D crack propagation. *Struct. Eng. Mech.* 2001; 12(5):475-490.
- Aymone JLF, Bittencourt E and Creus GJ. Simulation of 3D metal forming using an arbitrary Lagrangian-Eulerian finite element method. *J. Mater. Process. Technol.* 2001; 110(2): 218-232.
- Baatjens FFT. An U-ALE formulation of 3-D unsteady viscoelastic flow. *Int. J. Numer. Methods Eng.* 1993; 36(7): 1115-1143.
- Batina JT. Implicit flux-split Euler schemes for unsteady aerodynamic analysis involving unstructured dynamic meshes. *AIAA J.* 1991; 29(11):1836-1843.
- Belytschko T and Kennedy JM. Computer methods for subassembly simulation. *Nucl. Eng. Des.* 1978; 49:17-38.
- Belytschko T and Liu WK. Computer methods for transient fluid-structure analysis of nuclear reactors. *Nucl. Saf.* 1985; 26(1):14-31.
- Belytschko T, Flanagan DP and Kennedy JM. Finite element methods with user-controlled meshes for fluid-structure interaction. *Comput. Methods Appl. Mech. Eng.* 1982; 33(1-3): 669-688.
- Belytschko T, Kennedy JM and Schoeberle DF. Quasi-Eulerian finite element formulation for fluid-structure interaction. *Proceedings of Joint ASME/CSME Pressure Vessels and Piping Conference*. ASME: New York, 1978; p. 13, ASME paper 78-PVP-60.
- Belytschko T, Kennedy JM and Schoeberle DF. Quasi-Eulerian finite element formulation for fluid-structure interaction. *J. Press. Vessel Technol.-Trans. ASME* 1980; 102:62-69.
- Belytschko T, Liu WK and Moran B. *Nonlinear Finite Elements for Continua and Structures*. Wiley: Chichester, 2000.
- Benson DJ. An efficient, accurate, simple ALE method for nonlinear finite element programs. *Comput. Methods Appl. Mech. Eng.* 1989; 72(3):305-350.
- Benson DJ. Vectorization techniques for explicit ALE calculations. An efficient, accurate, simple ALE method for nonlinear finite element programs. *Comput. Methods Appl. Mech. Eng.* 1992a; 96(3):303-328.
- Benson DJ. Computational methods in Lagrangian and Eulerian hydrocodes. *Comput. Meth. Appl. Mech. Eng.* 1992b; 99(2-3):235-394.
- Boman R and Ponthot JP. Finite elements for the lubricated contact between solids in metal forming processes. *Acta Metall. Sinica* 2000; 13(1):319-327.
- Boman R and Ponthot JP. Numerical simulation of lubricated contact in rolling processes. *J. Mater. Process. Technol.* 2002; 125-126:405-411.
- Bonet J and Wood RD. *Nonlinear Continuum Mechanics for Finite Element Analysis*. Cambridge University Press: Cambridge, 1997.
- Braess H and Wriggers P. Arbitrary Lagrangian-Eulerian finite element analysis of free surface flow. *Comput. Methods Appl. Mech. Eng.* 2000; 190(1-2):95-110.
- Casadei F and Halleux JP. An algorithm for permanent fluid-structure interaction in explicit transient dynamics. *Comput. Methods Appl. Mech. Eng.* 1995; 128(3-4):231-289.
- Casadei F and Potapov S. Permanent fluid-structure interaction with nonconforming interfaces in fast transient dynamics. *Comput. Methods Appl. Mech. Eng.* 2004; 193: to appear in the Special Issue on the ALE formulation.
- Casadei F and Sala A. Finite element and finite volume simulation of industrial fast transient fluid-structure interactions. In *Proceedings European Conference on Computational Mechanics - Solids, Structures and Coupled Problems in Engineering*, Wunderlich W (ed.), Lehrstuhl für Statik: Technische Universität München, 1999.
- Casadei F, Halleux JP, Sala A and Chilli F. Transient fluid-structure interaction algorithms for large industrial applications. *Comput. Methods Appl. Mech. Eng.* 2001; 190(24-25): 3081-3110.
- Castro-Diaz MJ, Bourouchaki H, George PL, Hecht F and Mohammadi B. Anisotropic adaptive mesh generation in two dimensions for CFD. In *Proceedings of the Third ECCOMAS Computational Fluid Dynamics Conference*, Paris, 9-13 September, 1996, 181-186.
- Cescutti JP, Wey B and Chenot JL. Finite element calculation of hot forging with continuous remeshing. In *Modelling of Metal Forming Processes, EUROMECH-233*, Chenot JL and Oñate E (eds), Sophia-Antipolis, 1988; 207-216.
- Chenot JL and Bellet M. The ALE method for the numerical simulation of material forming processes. In *Simulation of Materials Processing: Theory, Methods and Applications - NUMIFORM 95*, Shen SF and Dawson P (eds), Balkema: Ithaca, New York, 1995; 39-48.
- Donea J. Arbitrary Lagrangian-Eulerian finite element methods. In *Computational Methods for Transient Analysis*, Belytschko T and Hughes TJR (eds), North-Holland: Amsterdam, 1983; 474-516.
- Donea J and Huerta A. *Finite Element Methods for Flow Problems*. Wiley: Chichester, 2003.
- Donea J, Fasoli-Stella P and Giuliani S. Lagrangian and Eulerian finite element techniques for transient fluid-structure interaction problems. In *Trans. 4th Int. Conf. on Structural Mechanics in Reactor Technology*, Paper B1/2, San Francisco, 1977.
- Donea J, Giuliani S and Halleux JP. An arbitrary Lagrangian-Eulerian finite element method for transient dynamic fluid-structure interactions. *Comput. Methods Appl. Mech. Eng.* 1982; 33(1-3):689-723.
- Eriksson LE. Practical three-dimensional mesh generation using transfinite interpolation. *SIAM J. Sci. Statist. Comput.* 1985; 6(3):712-741.
- Farhat C, Geuzaine Ph and Grandmont C. The discrete geometric conservation law and the nonlinear stability of ALE schemes for the solution of flow problems on moving grids. *J. Comput. Phys.* 2001; 174(2):669-694.
- Fortin M, Vallet M-G, Dompiere J, Bourgault Y and Habashi WG. Anisotropic mesh adaptation: theory, validation and applications. In *Proceedings of the Third ECCOMAS Computational Fluid Dynamics Conference*, Paris, 9-13 September, 1996, 174-180.
- Frank RM and Lazarus RB. Mixed Eulerian-Lagrangian method. In *Methods in Computational Physics*, Vol. 3: *Fundamental Methods in Hydrodynamics*, Alder B, Fernbach S and Rotenberg M (eds), Academic Press: New York, 1964; 47-67.
- Gadala MS and Wang J. ALE formulation and its application in solid mechanics. *Comput. Methods Appl. Mech. Eng.* 1998; 167(1-2):33-55.
- Gadala MS and Wang J. Simulation of metal forming processes with finite element methods. *Int. J. Numer. Methods Eng.* 1999; 44(10):1397-1428.
- Gadala MS, Movahhedy MR and Wang J. On the mesh motion for ALE modeling of metal forming processes. *Finite Elem. Anal. Des.* 2002; 38(5):435-459.
- Ghosh S. Arbitrary Lagrangian-Eulerian finite element analysis of large deformation in contacting bodies. *Int. J. Numer. Methods Eng.* 1992; 33(9):1891-1925.
- Ghosh S and Kikuchi N. An arbitrary Lagrangian-Eulerian finite element method for large deformation analysis of elastic-viscoplastic solids. *Comput. Methods Appl. Mech. Eng.* 1991; 86(2):127-188.
- Ghosh S and Raju S. R-S adapted arbitrary Lagrangian Eulerian finite element method for metal-forming problems with strain localization. *Int. J. Numer. Methods Eng.* 1996; 39(19): 3247-3272.
- Giuliani S. An algorithm for continuous rezoning of the hydrodynamic grid in arbitrary Lagrangian-Eulerian computer codes. *Nucl. Eng. Des.* 1982; 72:205-212.
- Gordon WJ and Hall CH. Construction of curvilinear co-ordinate systems and applications to mesh generation. *Int. J. Numer. Methods Eng.* 1973; 7(4):461-477.
- Guillard H and Farhat C. On the significance of the geometric conservation law for flow computations on moving meshes. *Comput. Methods Appl. Mech. Eng.* 2000; 190(11-12):1467-1482.
- Habashi WG, Fortin M, Vallet M-G, Dompiere J, Bourgault Y and Ait-Ali-Yahia D. Anisotropic mesh adaptation: towards user-independent, mesh-independent and solver-independent CFD solutions. Part I: Theory. *Int. J. Numer. Methods Fluids* 2000; 32(6):725-744.
- Haber RB. A mixed Eulerian-Lagrangian displacement model for large-deformation analysis in solid mechanics. *Comput. Methods Appl. Mech. Eng.* 1984; 43(3):277-292.
- Haber R and Abel JF. Discrete transfinite mappings for the description and meshing of three-dimensional surfaces using interactive computer graphics. *Int. J. Numer. Methods Eng.* 1982; 18(1):41-66.
- Haber RB and Hariandja BH. Computational strategies for nonlinear and fracture mechanics problems. An Eulerian-Lagrangian finite element approach to large deformation frictional contact. *Comput. Struct.* 1985; 20(1-3):193-201.
- Hermansson J and Hansbo P. A variable diffusion method for mesh smoothing. *Commun. Numer. Methods Eng.* 2003; 19(11):897-908.
- Hirt CW, Amsden AA and Cook JL. An arbitrary Lagrangian-Eulerian computing method for all flow speeds. *J. Comput. Phys.* 1974; 14:227-253. Reprinted in *J. Comput. Phys.* 1997; 135(2):203-216.
- Hogge M and Ponthot JP. Metal forming analysis via Eulerian-Lagrangian FEM with adaptive mesh. *Latin Am. Res.* 1991; 21:217-224.
- Hu YK and Liu WK. ALE finite element formulation for ring rolling analysis. *Int. J. Numer. Methods Eng.* 1992; 33(6):1217-1236.
- Hu YK and Liu WK. An ALE hydrodynamic lubrication finite element method with application to strip rolling. *Int. J. Numer. Methods Eng.* 1993; 36(5):855-880.
- Hu YK and Liu WK. Finite element hydrodynamic friction model for metal forming. *Int. J. Numer. Methods Eng.* 1994; 37(23):4015-4037.
- Huerta A and Casadei F. New ALE applications in non-linear fast-transient solid dynamics. *Eng. Comput.* 1994; 11(4):317-345.
- Huerta A and Liu WK. Viscous flow structure interaction. *J. Press. Vessel Technol.-Trans. ASME* 1988a; 110(1):15-21.
- Huerta A and Liu WK. Viscous flow with large free-surface motion. *Comput. Methods Appl. Mech. Eng.* 1988b; 69(3): 277-324.
- Huerta A and Liu WK. ALE formulation for large boundary motion. In *Trans. 10th Int. Conf. Structural Mechanics in Reactor Technology*, Vol. B, Anaheim, 1989; 335-346.
- Huerta A and Liu WK. Large amplitude sloshing with submerged blocks. *J. Press. Vessel Technol.-Trans. ASME* 1990; 112:104-108.
- Huerta A, Rodríguez-Ferran A, Díez P and Sarraite J. Adaptive finite element strategies based on error assessment. *Int. J. Numer. Methods Eng.* 1999; 46(10):1803-1818.
- Huétink J, Vreede PT and van der Lugt J. Progress in mixed Eulerian-Lagrangian finite element simulation of forming processes. *Int. J. Numer. Methods Eng.* 1990; 30(8):1441-1457.
- Hughes TJR, Liu WK and Zimmermann TK. Lagrangian-Eulerian finite element formulation for incompressible viscous flows. *Comput. Methods Appl. Mech. Eng.* 1981; 29(3):329-349; Presented at the U.S.-Japan conference on Interdisciplinary Finite Element Analysis, Cornell University, August 7-11, 1978.
- Koh HM and Haber RB. Elastodynamic formulation of the Eulerian-Lagrangian kinematic description. *J. Appl. Mech.-Trans. ASME* 1986; 53(4):839-845.
- Koh HM, Lee HS and Haber RB. Dynamic crack propagation analysis using Eulerian-Lagrangian kinematic descriptions. *Comput. Mech.* 1988; 3:141-155.

- Koobus B and Farhat C. Second-order accurate and geometrically conservative implicit schemes for flow computations on unstructured dynamic meshes. *Comput. Methods Appl. Mech. Eng.* 1999; **170**(1-2):103-129.
- Laursen TA. *Computational Contact and Impact Mechanics*. Springer; Berlin, 2002.
- Lesoinne M and Farhat C. Geometric conservation laws for flow problems with moving boundaries and deformable meshes, and their impact on aerodynamic computations. *Comput. Methods Appl. Mech. Eng.* 1996; **134**(1-2):71-90.
- Le Tallec P and Martin C. A nonlinear elasticity model for structured mesh adaptation. In *Proceedings of the Third ECCOMAS Computational Fluid Dynamics Conference*, Paris, 1996; 174-180.
- Le Tallec P and Mouro J. Fluid-structure interaction with large structural displacements. *Comput. Methods Appl. Mech. Eng.* 2001; **190**(24-25):3039-3067.
- LeVeque RJ. *Numerical Methods for Conservation Laws*. Lectures in Mathematics, ETH Zürich. Birkhäuser Verlag; Basel, 1990.
- Liu WK and Chang HG. Efficient computational procedures for long-time duration fluid-structure interaction problems. *J. Press. Vessel Technol.-Trans. ASME* 1984; **106**:317-322.
- Liu WK and Chang HG. A method of computation for fluid structure interaction. *Comput. Struct.* 1985; **20**(1-3):311-320.
- Liu WK and Gvildys J. Fluid-structure interaction of tanks with an eccentric core barrel. *Comput. Methods Appl. Mech. Eng.* 1986; **58**(1):51-77.
- Liu WK, Belytschko T and Chang H. An arbitrary Lagrangian-Eulerian finite element method for path-dependent materials. *Comput. Methods Appl. Mech. Eng.* 1986; **58**(2):227-245.
- Liu WK, Chang H, Chen JS and Belytschko T. Arbitrary Lagrangian-Eulerian Petrov-Galerkin finite elements for nonlinear continua. *Comput. Methods Appl. Mech. Eng.* 1988; **68**(3):259-310.
- Liu WK, Chen JS, Belytschko T and Zhang YF. Adaptive ALE finite elements with particular reference to external work rate on frictional interface. *Comput. Methods Appl. Mech. Eng.* 1991; **93**(2):189-216.
- Löhner R and Yang C. Improved ALE mesh velocities for moving bodies. *Commun. Numer. Methods Eng.* 1996; **12**(10):599-608.
- Malvern LW. *Introduction to the Mechanics of a Continuous Medium*. Prentice-Hall; Englewood Cliffs, 1969.
- Martinet F and Chabrand P. Application of ALE finite element method to a lubricated friction model in sheet metal forming. *Int. J. Solids Struct.* 2000; **37**(29):4005-4031.
- Müller J-D. Anisotropic adaptation and multigrid for hybrid grids. *Int. J. Numer. Methods Fluids* 2002; **40**(3-4):445-455.
- Noh WF. CEL: A time-dependent two-space dimensional coupled Eulerian-Lagrangian code. In *Methods in Computational Physics*, Alder B, Fernbach S and Rotenberg M (eds), vol. 3. Academic Press; New York, 1964; 117-179.
- Nomura T and Hughes TJR. An arbitrary Lagrangian-Eulerian finite element method for interaction of fluid and a rigid body. *Comput. Methods Appl. Mech. Eng.* 1992; **95**(1):115-138.
- Pérez-Foguet A, Rodríguez-Ferran A and Huerta A. Efficient and accurate approach for powder compaction problems. *Comput. Mech.* 2003; **30**(3):220-234.
- Pijaudier-Cabot G, Bodé L and Huerta A. Arbitrary Lagrangian-Eulerian finite element analysis of strain localization in transient problems. *Int. J. Numer. Methods Eng.* 1995; **38**(24):4171-4191.
- Ponthot JP and Belytschko T. Arbitrary Lagrangian-Eulerian formulation for element-free Galerkin method. *Comput. Methods Appl. Mech. Eng.* 1998; **152**(1-2):19-46.
- Ponthot JP and Hogge M. The use of the Eulerian-Lagrangian FEM in metal forming applications including contact and adaptive mesh. In *Advances in Finite Deformation Problems in Material Processing*, Chandra N and Reddy JN (eds). ASME Winter Annual Meeting, ASME: AMD-125, ASME (American Society of Mechanical Engineers); Atlanta, 1991; 44-64.
- Pracht WE. Calculating three-dimensional fluid flows at all flow speeds with an Eulerian-Lagrangian computing mesh. *J. Comput. Phys.* 1975; **17**:152-159.
- Ramaswamy B and Kawahara M. Arbitrary Lagrangian-Eulerian finite element method for unsteady, convective, incompressible viscous free surface fluid flow. *Int. J. Numer. Methods Fluids* 1987; **7**(10):1053-1075.
- Rodríguez-Ferran A, Casadei F and Huerta A. ALE stress update for transient and quasistatic processes. *Int. J. Numer. Methods Eng.* 1998; **43**(2):241-262.
- Rodríguez-Ferran A, Pérez-Foguet A and Huerta A. Arbitrary Lagrangian-Eulerian (ALE) formulation for hyperelastoplasticity. *Int. J. Numer. Methods Eng.* 2002; **53**(8):1831-1851.
- Sarrate J and Huerta A. An improved algorithm to smooth graded quadrilateral meshes preserving the prescribed element size. *Commun. Numer. Methods Eng.* 2001; **17**(2):89-99.
- Sarrate J, Huerta A and Donea J. Arbitrary Lagrangian-Eulerian formulation for fluid rigid-body interaction. *Comput. Methods Appl. Mech. Eng.* 2001; **190**(24-25):3171-3188.
- Schreurs PJG, Veldpaus FE and Brekelmans WAM. Simulation of forming processes using the Arbitrary Eulerian-Lagrangian formulation. *Comput. Methods Appl. Mech. Eng.* 1986; **58**(1):19-36.
- Souti M and Zolesio JP. Arbitrary Lagrangian-Eulerian and free-surface methods in fluid mechanics. *Comput. Methods Appl. Mech. Eng.* 2001; **191**(3-5):451-466.
- Smith RW. AUSM(ALE): a geometrically conservative Arbitrary Lagrangian-Eulerian flux splitting scheme. *J. Comput. Phys.* 1999; **150**(1):268-286.
- Trepanier JY, Reggio M, Parasciovo M and Camarero R. Unsteady Euler solutions for arbitrarily moving bodies and boundaries. *AIAA J.* 1993; **31**(10):1869-1876.
- Trulio JC. *Theory and Structure of the AFTON Codes*. Report AFWL-TR-66-19, Air Force Weapons Laboratory: Kirtland Air Force Base, 1966.
- van Haaren MJ, Stoker HC, van den Boogaard AH and Huétink J. The ALE method with triangular elements: Direct convection of integration point values. *Int. J. Numer. Methods Eng.* 2000; **49**(5):697-720.
- Winslow AM. *Equipotential Zoning of Two-dimensional Meshes*. Report UCRL-7312, University of California, Lawrence Livermore Laboratory; Livermore, California, 1963.
- Wriggers P. *Computational Contact Mechanics*. Wiley; Chichester, 2002.
- Yamada T and Kikuchi F. An arbitrary Lagrangian-Eulerian finite element method for incompressible hyperelasticity. *Comput. Methods Appl. Mech. Eng.* 1993; **102**(2):149-177.
- Zhang Q and Hisada T. Analysis of fluid-structure interaction problems with structural buckling and large domain changes by ALE finite element method. *Comput. Methods Appl. Mech. Eng.* 2001; **190**(48):6341-6357.
- Zhong ZH. *Finite Element Procedures for Contact-Impact Problems*. Oxford University Press; Oxford, 1993.

Chapter 15

Finite Volume Methods: Foundation and Analysis

Timothy Barth¹ and Mario Ohlberger²

¹NASA Ames Research Center, Moffett Field, CA, USA
²Freiburg University, Freiburg, Germany and University of Maryland, College Park, MD, USA

| | |
|--|-----|
| 1 Introduction: Scalar Nonlinear Conservation Laws | 439 |
| 2 Finite Volume (FV) Methods for Nonlinear Conservation Laws | 442 |
| 3 Higher-order Accurate FV Generalizations | 450 |
| 4 Further Advanced Topics | 464 |
| 5 Concluding Remarks | 470 |
| 6 Related Chapters | 470 |
| References | 470 |

1 INTRODUCTION: SCALAR NONLINEAR CONSERVATION LAWS

Many problems arising in science and engineering lead to the study of nonlinear hyperbolic conservation laws. Some examples include fluid mechanics, meteorology, electromagnetics, semi conductor device simulation, and numerous models of biological processes. As a prototype conservation law, consider the Cauchy initial value problem

$$\partial_t u + \nabla \cdot f(u) = 0 \quad \text{in } \mathbb{R}^d \times \mathbb{R}^+ \tag{1a}$$

$$u(x, 0) = u_0(x) \quad \text{in } \mathbb{R}^d \tag{1b}$$

Here $u(x, t): \mathbb{R}^d \times \mathbb{R}^+ \rightarrow \mathbb{R}$ denotes the dependent solution variable, $f(u) \in C^1(\mathbb{R})$ denotes the flux function, and $u_0(x): \mathbb{R}^d \rightarrow \mathbb{R}$ the initial data.

Encyclopedia of Computational Mechanics, Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes. Volume 1: *Fundamentals*. © 2004 John Wiley & Sons, Ltd. ISBN: 0-470-84699-2.

The function u is a *classical solution* of the scalar initial value problem if $u \in C^1(\mathbb{R}^d \times \mathbb{R}^+)$ satisfies (1a, 1b) pointwise. An essential feature of nonlinear conservation laws is that, in general, gradients of u blow up in finite time, even when the initial data u_0 is arbitrarily smooth. Beyond some critical time t_0 classical solutions of (1a, 1b) do not exist. This behavior will be demonstrated shortly using the method of characteristics. By introducing the notion of weak solutions of (1a, 1b) together with an entropy condition, it then becomes possible to define a class of solutions where existence and uniqueness is guaranteed for times greater than t_0 . These are precisely the solutions that are numerically sought in the finite volume method.

1.1 The method of characteristics

Let u be a classical solution of (1a) subject to initial data (1b). Further, define the vector

$$a(u) = f'(u) = (f'_1(u), \dots, f'_d(u))^T$$

A characteristic Γ_ξ is a curve $(x(t), t)$ such that

$$x'(t) = a(u(x(t), t)) \quad \text{for } t > 0$$

$$x(0) = \xi$$

Since u is assumed to be a classical solution, it is readily verified that

$$\begin{aligned} \frac{d}{dt}u(x(t), t) &= \partial_t u + x'(t) \nabla u \\ &= \partial_t u + a(u) \nabla u = \partial_t u + \nabla \cdot f(u) = 0 \end{aligned}$$

Therefore, u is constant along a characteristic curve and Γ_ξ is a straight line since

$$\begin{aligned} x'(t) &= a(u(x(t), t)) = a(u(x(0), 0)) \\ &= a(u(\xi, 0)) = a(u_0(\xi)) = \text{const} \end{aligned}$$

In particular, $x(t)$ is given by

$$x(t) = \xi + ta(u_0(\xi)) \quad (2)$$

This important property may be used to construct classical solutions. If x and t are fixed and ξ determined as a solution of (2), then

$$u(x, t) = u_0(\xi)$$

This procedure is the basis of the so-called method of characteristics. On the other hand, this construction shows that the intersection of any two straight characteristic lines leads to a contradiction in the definition of $u(x, t)$. Thus, classical solutions can only exist up to the first time t_0 at which any two characteristics intersect.

1.2 Weak solutions

Since, in general, classical solutions only exist for a finite time t_0 , it is necessary to introduce the notion of weak solutions that are well defined for times $t > t_0$.

Definition 1 (Weak solution) Let $u_0 \in L^\infty(\mathbb{R}^d)$. Then, u is a weak solution of (1a, 1b) if $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}^+)$ and (1a, 1b) hold in the distributional sense, that is,

$$\begin{aligned} \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} (u \partial_t \phi + f(u) \cdot \nabla \phi) \, dt \, dx \\ + \int_{\mathbb{R}^d} u_0 \phi(x, 0) \, dx = 0 \quad \text{for all } \phi \in C_0^\infty(\mathbb{R}^d \times \mathbb{R}^+) \end{aligned} \quad (3)$$

Note that classical solutions are weak solutions and weak solutions that lie in $C^1(\mathbb{R}^d \times \mathbb{R}^+)$ satisfy (1a, 1b) in the classical sense.

It can be shown (see Kruzkov, 1970; Oleinik, 1963) that there always exists at least one weak solution to (1a, 1b) if the flux function f is at least Lipschitz continuous. Nevertheless, the class of weak solutions is too large to ensure uniqueness of solutions. An important class of solutions are piecewise classical solutions with discontinuities separating the smooth regions. The following lemma gives a necessary and sufficient condition imposed on these discontinuities such that the solution is a weak solution; see, for example, Godlewski and Raviart (1991) and Kröner (1997).

Later a simple example is given where infinitely many weak solutions exist.

Lemma 1 (Rankine–Hugoniot jump condition) Assume that $\mathbb{R}^d \times \mathbb{R}^+$ is separated by a smooth hypersurface S into two parts Ω_+ and Ω_- . Furthermore, assume u is a C^1 -function on Ω_+ and Ω_- , respectively. Then, u is a weak solution of (1a, 1b) if and only if the following two conditions hold:

- u is a classical solution in Ω_+ and Ω_- .
- u satisfies the Rankine–Hugoniot jump condition, that is,

$$[u]s = [f(u)] \cdot \nu \quad \text{on } S \quad (4)$$

Here, $(\nu, -s)^T$ denotes a unit normal vector for the hypersurface S and $[u]$ denotes the jump in u across the hypersurface S .

In one space dimension, it may be assumed that S is parameterized by $(\sigma(t), t)$ such that $s = \sigma'(t)$ and $\nu = 1$. The Rankine–Hugoniot jump condition then reduces to

$$s = \frac{[f(u)]}{[u]} \quad \text{on } S \quad (5)$$

Example 1 (Non uniqueness of weak solutions) Consider the one-dimensional Burgers' equation, $f(u) = u^2/2$, with Riemann data: $u_0(x) = u_l$ for $x < 0$ and $u_0(x) = u_r$ for $x \geq 0$. Then, for any $a \geq \max(u_l, -u_r)$ a function u given by

$$u(x, t) = \begin{cases} u_l, & x < s_1 t \\ -a, & s_1 t < x < 0 \\ a, & 0 < x < s_2 t \\ u_r, & s_2 t < x \end{cases} \quad (6)$$

is a weak solution if $s_1 = (u_l - a)/2$ and $s_2 = (a + u_r)/2$. This is easily checked since u is piecewise constant and satisfies the Rankine–Hugoniot jump condition. This elucidates a one-parameter family of weak solutions. In fact, there is also a classical solution whenever $u_l \leq u_r$. In this case, the characteristics do not intersect and the method of characteristics yields the classical solution

$$u(x, t) = \begin{cases} u_l, & x < u_l t \\ x/t, & u_l t < x < u_r t \\ u_r, & u_r t < x \end{cases} \quad (7)$$

This solution is the unique classical solution but not the unique weak solution. Consequently, additional conditions must be introduced in order to single out one solution within the class of weak solutions. These additional conditions give rise to the notion of a unique entropy weak solution.

1.3 Entropy weak solutions and vanishing viscosity

In order to introduce the notion of entropy weak solutions, it is useful to first demonstrate that there is a class of additional conservation laws for any classical solution of (1a). Let u be a classical solution and $\eta: \mathbb{R} \rightarrow \mathbb{R}$ a smooth function. Multiplying (1a) by $\eta'(u)$, one obtains

$$0 = \eta'(u) \partial_t u + \eta'(u) \nabla \cdot f(u) = \partial_t \eta(u) + \nabla \cdot F(u) \quad (8)$$

where F is any primitive of $\eta' f'$. This reveals that for a classical solution u , the quantity $\eta(u)$, henceforth called an entropy function, is a conserved quantity.

Definition 2 (Entropy–entropy flux pair) Let $\eta: \mathbb{R} \rightarrow \mathbb{R}$ be a smooth convex function and $F: \mathbb{R} \rightarrow \mathbb{R}$ a smooth function such that

$$F' = \eta' f' \quad (9)$$

in (8). Then (η, F) is called an entropy–entropy flux pair or more simply an entropy pair for the equation (1a).

Note 1 (Kruzkov entropies) The family of smooth convex entropies η may be equivalently replaced by the nonsmooth family of so-called Kruzkov entropies, that is, $\eta_\kappa(u) = |u - \kappa|$ for all $\kappa \in \mathbb{R}$ (see Kröner, 1997).

Unfortunately, the relation (8) can not be fulfilled for weak solutions in general, as it would lead to additional jump conditions that would contradict the Rankine–Hugoniot jump condition lemma. Rather, a weak solution may satisfy the relation (8) in the distributional sense with inequality. To see that this concept of entropy effectively selects a unique, physically relevant solution among all weak solutions, consider the viscosity perturbed equation

$$\partial_t u_\epsilon + \nabla \cdot f(u_\epsilon) = \epsilon \Delta u_\epsilon \quad (10)$$

with $\epsilon > 0$. For this parabolic problem, it may be assumed that a unique smooth solution u_ϵ exists. Multiplying by η' and rearranging terms yields the additional equation

$$\partial_t \eta(u_\epsilon) + \nabla \cdot F(u_\epsilon) = \epsilon \Delta \eta(u_\epsilon) - \epsilon \eta''(u_\epsilon) |\nabla u|^2$$

Furthermore, since η is assumed convex ($\eta'' \geq 0$), the following inequality is obtained

$$\partial_t \eta(u_\epsilon) + \nabla \cdot F(u_\epsilon) \leq \epsilon \Delta \eta(u_\epsilon)$$

Taking the limit $\epsilon \rightarrow 0$ establishes (see Málek, Nečas, Rokyta and Růžička, 1996) that u_ϵ converges towards

some u a.e. in $\mathbb{R}^d \times \mathbb{R}^+$ where u is a weak solution of (1a, 1b) and satisfies the entropy condition

$$\partial_t \eta(u) + \nabla \cdot F(u) \leq 0 \quad (11)$$

in the sense of distributions on $\mathbb{R}^d \times \mathbb{R}^+$.

By this procedure, a unique weak solution has been identified as the limit of the approximating sequence u_ϵ . The obtained solution u is called the vanishing viscosity weak solution of (1a, 1b). Motivated by the entropy inequality (11) of the vanishing viscosity solution, it is now possible to introduce the notion of entropy weak solutions. This notion is weak enough for the existence and strong enough for the uniqueness of solutions to (1a, 1b).

Definition 3 (Entropy weak solution) Let u be a weak solution of (1a, 1b). Then, u is called an entropy weak solution if u satisfies for all entropy pairs (η, F)

$$\begin{aligned} \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} (\eta(u) \partial_t \phi + F(u) \cdot \nabla \phi) \, dt \, dx \\ + \int_{\mathbb{R}^d} \eta(u_0) \phi(x, 0) \, dx \geq 0 \end{aligned}$$

for all $\phi \in C_0^\infty(\mathbb{R}^d \times \mathbb{R}^+, \mathbb{R}^+)$ (12)

From the vanishing viscosity method, it is known that entropy weak solutions exist. The following L^1 contraction principle guarantees that entropy solutions are uniquely defined; see Kruzkov (1970).

Theorem 1 (L^1 -contraction principle) Let u and v be two entropy weak solutions of (1a, 1b) with respect to initial data u_0 and v_0 . Then, the following L^1 -contraction principle holds

$$\|u(\cdot, t) - v(\cdot, t)\|_{L^1(\mathbb{R}^d)} \leq \|u_0 - v_0\|_{L^1(\mathbb{R}^d)} \quad (13)$$

for almost every $t > 0$.

This principle demonstrates a continuous dependence of the solution on the initial data and consequently the uniqueness of entropy weak solutions. Finally, note that an analog of the Rankine–Hugoniot condition exists (with inequality) in terms of the entropy pair for all entropy weak solutions

$$[\eta(u)]s \geq [F(u)] \cdot \nu \quad \text{on } S \quad (14)$$

1.4 Measure-valued or entropy process solutions

The numerical analysis of conservation laws requires an even weaker formulation of solutions to (1a, 1b). For instance, the convergence analysis of finite volume schemes

makes it necessary to introduce so-called measure-valued or entropy process solutions; see DiPerna (1985) and Eymard, Gallu  t and Herbin (2000).

Definition 4 (Entropy process solution) A function $\mu(x, t, \alpha) \in L^\infty(\mathbb{R}^d \times \mathbb{R}^+ \times (0, 1))$ is called an entropy process solution of (1a, 1b) if μ satisfies for all entropy pairs (η, F)

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^+} \int_0^1 \eta(\mu) \partial_t (\phi + F(\mu) \cdot \nabla \phi) dx dt dx + \int_{\mathbb{R}^d} \eta(u_0) \phi(x, 0) dx \geq 0$$

for all $\phi \in C_0^\infty(\mathbb{R}^d \times \mathbb{R}^+, \mathbb{R}^+)$

The most important property of such entropy process solutions is the following uniqueness and regularity result (see Eymard, Gallu  t and Herbin, 2000, Theorem 6.3).

Theorem 2 (Uniqueness of entropy process solutions) Let $u_0 \in L^\infty(\mathbb{R}^d)$ and $f \in C^1(\mathbb{R})$. The entropy process solution μ of problem (1a, 1b) is unique. Moreover, there exists a function $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}^+)$ such that $u(x, t) = \mu(x, t, \alpha)$ a.e. for $(x, t, \alpha) \in \mathbb{R}^d \times \mathbb{R}^+ \times (0, 1)$ and u is the unique entropy weak solution of (1a, 1b).

2 FINITE VOLUME (FV) METHODS FOR NONLINEAR CONSERVATION LAWS

In the finite volume method, the computational domain, $\Omega \subset \mathbb{R}^d$, is first tessellated into a collection of nonoverlapping control volumes that completely cover the domain. Notationally, let \mathcal{T} denote a tessellation of the domain Ω with control volumes $T \in \mathcal{T}$ such that $\bigcup_{T \in \mathcal{T}} T = \Omega$. Let h_T denote a length scale associated with each control volume T , for example, $h_T = \text{diam}(T)$. For two distinct control volumes T_i and T_j in \mathcal{T} , the intersection is either an oriented edge (2-D) or face (3-D) e_{ij} with oriented normal ν_{ij} or else a set of measure at most $d - 2$. In each control volume, an integral conservation law statement is then imposed.

Definition 5 (Integral conservation law) An integral conservation law asserts that the rate of change of the total amount of a substance with density u in a fixed control volume T is equal to the total flux of the substance through the boundary ∂T

$$\frac{d}{dt} \int_T u dx + \int_{\partial T} f(u) \cdot \nu = 0 \quad (15)$$

This integral conservation law statement is readily obtained upon spatial integration of the divergence equation (1a) in

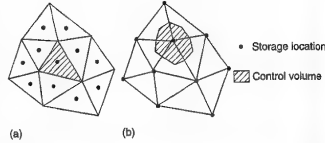


Figure 1. Control volume variants used in the finite volume method (a) cell-centered and (b) vertex-centered control volume tessellation.

the region T and application of the divergence theorem. The choice of control volume tessellation is flexible in the finite volume method. For example, Figure 1 depicts a 2-D triangle complex and two typical control volume tessellations (among many others) used in the finite volume method. In the cell-centered finite volume method shown in Figure 1(a), the triangles themselves serve as control volumes with solution unknowns (degrees of freedom) stored on a per triangle basis. In the vertex-centered finite volume method shown in Figure 1(b), control volumes are formed as a geometric dual to the triangle complex and solution unknowns stored on a per triangulation vertex basis.

2.1 Godunov finite volume discretizations

Fundamental to finite volume methods is the introduction of the control volume cell average for each $T_j \in \mathcal{T}$

$$u_j = \frac{1}{|T_j|} \int_{T_j} u dx \quad (16)$$

For stationary meshes, the finite volume method can be interpreted as producing an evolution equation for cell averages

$$\frac{d}{dt} \int_{T_j} u dx = |T_j| \frac{d}{dt} u_j \quad (17)$$

Godunov (1959) pursued this interpretation in the discretization of the gas dynamic equations by assuming piecewise constant solution representations in each control volume with value equal to the cell average. However, the use of piecewise constant representations renders the numerical solution multivalued at control volume interfaces thereby making the calculation of a single solution flux at these interfaces ambiguous. The second aspect of Godunov's scheme and subsequent variants was the idea of supplanting the true flux at interfaces by a numerical flux function, $g(u, v): \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, a Lipschitz continuous function of the two interface states u and v . A single

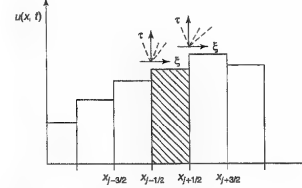


Figure 2. 1-D control volume, $T_j = [x_{j-1/2}, x_{j+1/2}]$, depicting Godunov's interface Riemann problem, $w_{j\pm 1/2}(\xi, \tau)$, from piecewise constant interface states.

unique numerical flux was then calculated from an exact or approximate local solution of the Riemann problem in gas dynamics posed at these interfaces. Figure 2 depicts a representative 1-D solution profile in Godunov's method. For a given control volume $T_j = [x_{j-1/2}, x_{j+1/2}]$, Riemann problems are solved at each interface $x_{j\pm 1/2}$. For example, at the interface $x_{j+1/2}$, the Riemann problem counterpart of (1a, 1b)

$$\partial_t w_{j+1/2}(\xi, \tau) + \partial_\xi f(w_{j+1/2}(\xi, \tau)) = 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^+$$

for $w_{j+1/2}(\xi, \tau) \in \mathbb{R}$ with initial data

$$w_{j+1/2}(\xi, 0) = \begin{cases} u_j & \text{if } \xi < 0 \\ u_{j+1} & \text{if } \xi > 0 \end{cases}$$

is solved either exactly or approximately. From this local solution, a single unique numerical flux at $x_{j+1/2}$ is computed from $g(u_j, u_{j+1}) = f(w_{j+1/2}(0, \tau > 0))$. This construction utilizes the fact that the solution of the Riemann problem at $\xi = 0$ is a constant for all time $\tau > 0$.

In higher-space dimensions, the flux integral appearing in (15) is similarly approximated by

$$\int_{\partial T_j} f(u) \cdot \nu dv \approx \sum_{e_{jk} \in \partial T_j} g_{jk}(u_j, u_k) \quad (18)$$

where the numerical flux is assumed to satisfy the properties:

- (Conservation) This property ensures that fluxes from adjacent control volumes sharing a mutual interface exactly cancel when summed. This is achieved if the numerical flux satisfies the identity

$$g_{jk}(u, v) = -g_{kj}(v, u) \quad (19a)$$

- (Consistency) Consistency is obtained if the numerical flux with identical state arguments reduces to the true total flux passing through e_{jk} of that same state, that is,

$$g_{jk}(u, u) = \int_{e_{jk}} f(u) \cdot \nu dv \quad (19b)$$

Combining (17) and (18) yields perhaps the simplest finite volume scheme in semidiscrete form. Let V_h^0 denote the space of piecewise constants, that is,

$$V_h^0 = \{v \mid v|_T \in \chi(T), \quad \forall T \in \mathcal{T}\} \quad (20)$$

with $\chi(T)$ a characteristic function in the control volume T .

Definition 6 (Semidiscrete finite volume method) The semidiscrete finite volume approximation of (1a, 1b) utilizing continuous in time solution representation, $t \in [0, \infty)$, and piecewise constant solution representation in space, $u_h(t) \in V_h^0$, such that

$$u_j(t) = \frac{1}{|T_j|} \int_{T_j} u_h(x, t) dx$$

with initial data

$$u_j(0) = \frac{1}{|T_j|} \int_{T_j} u_0(x) dx$$

and numerical flux function $g_{jk}(u_j, u_k)$ is given by the following system of ordinary differential equations

$$\frac{d}{dt} u_j + \frac{1}{|T_j|} \sum_{e_{jk} \in \partial T_j} g_{jk}(u_j, u_k) = 0, \quad \forall T_j \in \mathcal{T} \quad (21)$$

This system of ordinary differential equations can be marched forward using a variety of explicit and implicit time integration methods. Let u_j^n denote a numerical approximation of the cell average solution in the control volume T_j at time $t^n \equiv n \Delta t$. A particularly simple time integration method is the forward Euler scheme

$$\frac{d}{dt} u_j \approx \frac{u_j^{n+1} - u_j^n}{\Delta t}$$

thus producing a fully discrete finite volume form.

Definition 7 (Fully discrete finite volume method) The fully discrete finite volume approximation of (1a, 1b) for the time slab interval $[t^n, t^{n+1} + \Delta t]$ utilizing Euler explicit time

advancement and piecewise constant solution representation in space, $u_h^n \in V_h^n$, such that

$$u_j^n = \frac{1}{|T_j|} \int_{T_j} u_h^n(x) dx$$

with initial data for $n = 0$

$$u_j^0 = \frac{1}{|T_j|} \int_{T_j} u_0(x) dx$$

and numerical flux function $g_{jk}(u_j^n, u_k^n)$ is given by the following fully discrete system

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{|T_j|} \sum_{v_{jk} \in \partial T_j} g_{jk}(u_j^n, u_k^n), \quad \forall T_j \in \mathcal{T} \quad (22)$$

Once space-time maximum principle properties of this fully discrete form are ascertained, Section 4.1 shows how higher-order accurate time integration schemes can be constructed that preserve these properties, albeit with a different time step restriction.

2.1.1 Monotone schemes

Unfortunately, the numerical flux conditions (19a) and (19b) are insufficient to guarantee convergence to entropy satisfying weak solutions (12) and additional numerical flux restrictions are necessary. To address this deficiency, Harten, Hyman and Lax (1976) provide the following result concerning the convergence of the fully discrete one-dimensional scheme to weak entropy satisfying solutions.

Theorem 3 (Monotone schemes and weak solutions) Consider a 1-D finite volume discretization of (1a, 1b) with $2k+1$ stencil on a uniformly spaced mesh in both time and space with corresponding mesh spacing parameters Δt and Δx

$$\begin{aligned} u_j^{n+1} &= H_j(u_{j+2k}, \dots, u_{j+1}, u_j, \dots, u_{j-k}) \\ &= u_j^n - \frac{\Delta t}{\Delta x} (g_{j+1/2} - g_{j-1/2}) \end{aligned} \quad (23)$$

and consistent numerical flux of the form

$$g_{j+1/2} = g(u_{j+2k}, \dots, u_{j+1}, u_j, \dots, u_{j-k+1})$$

that is monotone in the sense

$$\frac{\partial H_j}{\partial u_{j+i}} \geq 0, \quad \forall |i| \leq k \quad (24)$$

Assume that u_j^n converges boundedly almost everywhere to some function $u(x, t)$, then as Δt and Δx tend to zero with

$\Delta t/\Delta x = \text{constant}$, this limit function $u(x, t)$ is an entropy satisfying weak solution of (1a, 1b).

Note that this theorem assumes convergence in the limit, which was later proven to be the case in multidimensions by Crandall and Majda (1980).

The monotonicity condition (24) motivates the introduction of Lipschitz continuous monotone fluxes satisfying

$$\frac{\partial g_{j+1/2}}{\partial u_l} \geq 0 \quad \text{if } l = j \quad (25a)$$

$$\frac{\partial g_{j+1/2}}{\partial u_l} \leq 0 \quad \text{if } l \neq j \quad (25b)$$

together with a CFL (Courant-Friedrichs-Levy) like condition

$$1 - \frac{\Delta t}{\Delta x} \left(\frac{\partial g_{j+1/2}}{\partial u_j} - \frac{\partial g_{j-1/2}}{\partial u_j} \right) \geq 0$$

so that (24) is satisfied. Some examples of monotone fluxes for (1a) include

- (Godunov flux)

$$g_{j+1/2}^G = \begin{cases} \min_{u \in [u_j, u_{j+1}]} f(u) & \text{if } u_j < u_{j+1} \\ \max_{u \in [u_j, u_{j+1}]} f(u) & \text{if } u_j > u_{j+1} \end{cases} \quad (26)$$

- (Lax-Friedrichs flux)

$$\begin{aligned} g_{j+1/2}^{LF} &= \frac{1}{2} (f(u_j) + f(u_{j+1})) \\ &\quad - \frac{1}{2} \sup_{u \in [u_j, u_{j+1}]} |f'(u)| (u_{j+1} - u_j) \end{aligned} \quad (27)$$

2.1.2 E-flux schemes

Another class of monotone numerical fluxes arising frequently in analysis was introduced by Osher (1984). These fluxes are called E-fluxes, $g_{j+1/2}^E = g^E(u_{j+2k}, \dots, u_{j+1}, u_j, \dots, u_{j-k+1})$, due to the relationship to Oleinik's well-known E-condition which characterizes entropy satisfying discontinuities. E-fluxes satisfy the inequality

$$\frac{g_{j+1/2}^E - f(u)}{u_{j+1} - u_j} \leq 0, \quad \forall u \in [u_j, u_{j+1}] \quad (28)$$

E-fluxes can be characterized by their relationship to Godunov's flux. Specifically, E-fluxes are precisely those fluxes such that

$$g_{j+1/2}^E \leq g_{j+1/2}^G \quad \text{if } u_{j+1} < u_j \quad (29a)$$

$$g_{j+1/2}^E \geq g_{j+1/2}^G \quad \text{if } u_{j+1} > u_j \quad (29b)$$

Viewed another way, note that any numerical flux can be written in the form

$$g_{j+1/2} = \frac{1}{2} (f(u_j) + f(u_{j+1})) - \frac{1}{2} Q(u_{j+1} - u_j) \quad (30)$$

where $Q(\cdot)$ denotes a viscosity for the scheme. When written in this form, E-fluxes are those fluxes that contribute at least as much viscosity as Godunov's flux, that is,

$$Q_{j+1/2}^G \leq Q_{j+1/2} \quad (31)$$

The most prominent E-flux is the Enquist-Osher flux

$$g_{j+1/2}^{EO} = \frac{1}{2} (f(u_j) - f(u_{j+1})) - \frac{1}{2} \int_{u_j}^{u_{j+1}} |f'(s)| ds \quad (32)$$

although other fluxes such as certain forms of Roe's flux with entropy fix fall into this category. From (29a, 29b), the monotone fluxes of Godunov $g_{j+1/2}^G$ and Lax-Friedrichs $g_{j+1/2}^{LF}$ are also E-fluxes.

2.2 Stability, convergence, and error estimates

Several stability results are provided here that originate from discrete maximum principle analysis and are straightforwardly stated in multidimensions and on general unstructured meshes. In presenting results concerning convergence and error estimates, a notable difference arises between one and several space dimensions. This is due to the lack of a BV bound on the approximate solution in multidimensions. Thus, before considering convergence and error estimates for finite volume methods, stability results are presented first together with some a priori bounds on the approximate solution.

2.2.1 Discrete maximum principles and stability

A compelling motivation for the use of monotone fluxes in the finite volume schemes (21) and (22) is the obtention of discrete maximum principles in the resulting numerical solution of nonlinear conservation laws (1a). A standard analysis technique is to first construct local discrete maximum principles which can then be applied successively to obtain global maximum principles and stability results.

The first result concerns the boundedness of local extrema in time for semidiscrete finite volume schemes that can be written in nonnegative coefficient form.

Theorem 4 (LED property) The semidiscrete scheme for each $T_j \in \mathcal{T}$

$$\frac{du_j}{dt} = \frac{1}{|T_j|} \sum_{v_{jk} \in \partial T_j} C_{jk}(u_k)(u_k - u_j) \quad (33)$$

is local extremum diminishing (LED), that is, local maxima are nonincreasing and local minima are nondecreasing, if

$$C_{jk}(u_k) \geq 0, \quad \forall v_{jk} \in \partial T_j \quad (34)$$

Rewriting the semidiscrete finite volume scheme (21) in the following equivalent forms

$$\begin{aligned} \frac{du_j}{dt} &= -\frac{1}{|T_j|} \sum_{v_{jk} \in \partial T_j} g_{jk}(u_j, u_k) \\ &= -\frac{1}{|T_j|} \sum_{v_{jk} \in \partial T_j} \frac{g_{jk}(u_j, u_k) - f(u_j) \cdot v_{jk} (u_k - u_j)}{u_k - u_j} \\ &= -\frac{1}{|T_j|} \sum_{v_{jk} \in \partial T_j} \frac{\partial g_{jk}}{\partial u_k}(u_j, u_k)(u_k - u_j) \end{aligned} \quad (35)$$

for appropriately chosen $\tilde{u}_{jk} \in [u_j, u_k]$ reveals that the monotone flux condition (25a, 25b) is a sufficient condition for the semidiscrete scheme to be LED. To obtain local space-time maximum principle results for the fully discrete discretization (22) requires the introduction of an additional CFL-like condition for nonnegativity of coefficients in space-time.

Theorem 5 (Local space-time discrete maximum principle) The fully discrete scheme for the time slab increment $[t^n, t^{n+1}]$ and each $T_j \in \mathcal{T}$

$$u_j^{n+1} = u_j^n + \frac{\Delta t}{|T_j|} \sum_{v_{jk} \in \partial T_j} C_{jk}(u_k^n)(u_k^n - u_j^n) \quad (36)$$

exhibits a local space-time discrete maximum principle for each $n = 0, 1, \dots$

$$\min_{v_{jk} \in \partial T_j} (u_k^n, u_j^n) \leq u_j^{n+1} \leq \max_{v_{jk} \in \partial T_j} (u_k^n, u_j^n) \quad (37)$$

if

$$C_{jk}(u_k^n) \geq 0, \quad \forall v_{jk} \in \partial T_j \quad (38)$$

and Δt is chosen such that the CFL-like condition is satisfied

$$1 - \frac{\Delta t}{|T_j|} \sum_{v_{jk} \in \partial T_j} C_{jk}(u_k^n) \geq 0 \quad (39)$$

Again noting that the flux terms in the fully discrete finite volume scheme (22) can be written in the form (35) reveals that the monotone flux conditions (25a, 25b) together with a local CFL-like condition obtained from (39) imply a local space-time discrete maximum principle. By successive application of Theorem 5, a global L^∞ -stability bound is

obtained for the scalar initial value problem (1a, 1b) in terms of initial data $u_0(x)$.

Theorem 6 (L^∞ -stability) Assume a fully discrete finite volume scheme (22) for the scalar initial value problem (1a, 1b) utilizing monotone fluxes satisfying a local CFL-like condition as given in Theorem 5 for each time slab increment $[t^n, t^{n+1}]$. Under these conditions, the finite volume scheme is L^∞ -stable and the following estimate holds:

$$\inf_{x \in \mathbb{R}^d} u_0(x) \leq u_j^n \leq \sup_{x \in \mathbb{R}^d} u_0(x), \quad \text{for all } (T_j, t^n) \in T \times [0, \tau] \quad (40)$$

Consider now steady state solutions, $u^{n+1} = u^n = u^*$, using a monotone flux in the fully discrete finite volume scheme (22). At steady state, nonnegativity of the coefficients $C(u_k)$ in (36) implies a discrete maximum principle.

Theorem 7 (Local discrete maximum principle in space) The fully discrete scheme (36) exhibits a local discrete maximum principle at steady state, u_k^* , for each $T_j \in T$

$$\min_{v_{e_k} \in \partial T_j} u_k^* \leq u_j^* \leq \max_{v_{e_k} \in \partial T_j} u_k^* \quad (41)$$

if

$$C_{jk}(u_k^*) \geq 0, \quad \forall e_{jk} \in \partial T_j$$

Once again, by virtue of (25a, 25b) and (28), the conditions for a local discrete maximum principle at steady state are fulfilled by monotone flux finite volume schemes (22). Global maximum principles for characteristic boundary value problems are readily obtained by successive application of the local maximum principle result.

The local maximum principles given in (37) and (41) preclude the introduction of spurious extrema and $\mathcal{O}(1)$ Gibbs-like oscillations that occur near solution discontinuities computed using many numerical methods (even in the presence of grid refinement). For this reason, discrete maximum principles of this type are a highly sought after design principle in the development of numerical schemes for nonlinear conservation laws.

2.2.2 Convergence results

The L^∞ -stability bound (40) is an essential ingredient in the proof of convergence of the fully discrete finite volume scheme (22). This bound permits the subtraction of a subsequence that converges against some limit in the L^∞ weak-star sense. The primary task that then remains is to identify this limit with the unique solution of the problem. So although L^∞ -stability is enough to ascertain

convergence of the scheme, stronger estimates are needed in order to derive convergence rates.

Let BV denote the space of functions with bounded variation, that is,

$$\text{BV} = \left\{ g \in L^1(\mathbb{R}^d) \mid |g|_{\text{BV}} < \infty \right\}$$

with

$$|g|_{\text{BV}} = \sup_{\substack{\varphi \in C_c^\infty(\mathbb{R}^d) \\ |\text{div}| \leq 1}} \int_{\mathbb{R}^d} g \nabla \cdot \varphi \, dx$$

From the theory of scalar conservation laws, it is known that, provided the initial data is in BV, the solution remains in BV for all times. Therefore, it is desirable to have an analog of this property for the approximate solution as well. Unfortunately, up to now, such a result is only rigorously proved in the one-dimensional case or in the case of tensor product Cartesian meshes in multiple space dimensions. In the general multidimensional case, the approximate solution can only be shown to fulfill some weaker estimate, which is thus called a weak BV estimate; see Vila (1994), Cockburn, Coquel and Lefloch (1994), and Eymard et al. (1998).

Theorem 8 (Weak BV estimate) Let T be a regular triangulation, and let J be a uniform partition of $[0, \tau]$ for example, $\Delta t^n = \Delta t$. Assume that there exists some $\alpha > 0$ such that $\alpha h^2 \leq |T_k|$, $\alpha |\partial T_k| \leq h$. For the time step Δt^n , assume the following CFL-like condition for a given $\xi \in (0, 1)$

$$\Delta t^n \leq \frac{(1 - \xi)\alpha^2 h}{L_g}$$

where L_g is the Lipschitz constant of the numerical flux function. Furthermore, let $u_0 \in L^\infty(\mathbb{R}^d) \cap BV(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Then, the numerical solution of the fully discrete discretization (22) fulfills the following estimate

$$\sum_n \Delta t \sum_j \chi_{j\ell} h |u_j^n - u_\ell^n| Q_{j\ell}(u_j^n, u_\ell^n) \leq K \sqrt{|T| B_{R+h}(0)} \sqrt{h} \quad (42)$$

where K only depends on α , L_g , ξ and the initial function u_0 . In this formula $Q_{j\ell}$ is defined as

$$Q_{j\ell}(u, v) \equiv \frac{2g_{j\ell}(u, v) - g_{j\ell}(u, u) - g_{j\ell}(v, v)}{u - v}$$

and $\chi_{j\ell}$ denotes the discrete cutoff function on $B_R(0) \subset \mathbb{R}^d$, that is,

$$\chi_{j\ell} = \begin{cases} 1, & \text{if } (T_j \cup T_\ell) \cap B_R(0) \neq \emptyset \\ 0, & \text{else} \end{cases}$$

Note that in the case of a strong BV estimate, the right-hand side of (42) would be $\mathcal{O}(h)$ instead of $\mathcal{O}(\sqrt{h})$.

Another important property of monotone finite volume schemes is that they preserve the L^1 -contraction property (see Theorem 1).

Theorem 9 (L^1 -contraction property and Lipschitz estimate in time) Let $u_h, v_h \in V_h^0$ be the approximate monotone finite volume solutions corresponding to initial data u_0, v_0 assuming that the CFL-like condition for stability has been fulfilled. Then the following discrete L^1 -contraction property holds

$$\|u_h(\cdot, t + \tau) - v_h(\cdot, t + \tau)\|_{L^1(\mathbb{R}^d)} \leq \|u_h(\cdot, t) - v_h(\cdot, t)\|_{L^1(\mathbb{R}^d)}$$

Furthermore, a discrete Lipschitz estimate in time is obtained

$$\sum_j |T_j| |u_j^{n+1} - u_j^n| \leq L_g \Delta t^n \sum_j \sum_\ell |e_{j\ell}| |u_j^n - u_\ell^n|$$

The principle ingredients of the convergence theory for scalar nonlinear conservation laws are compactness of the family of approximate solutions and the passage to the limit within the entropy inequality (12). In dealing with nonlinear equations, strong compactness is needed in order to pass to the limit in (12). In one space dimension, due to the BV estimate and the selection principle of Helly, strong compactness is ensured and the passage to the limit is summarized in the well-known Lax-Wendroff theorem; see Lax and Wendroff (1960).

Theorem 10 (Lax-Wendroff theorem) Let $(u_m)_{m \in \mathbb{N}}$ be a sequence of discrete solutions defined by the finite volume scheme in one space dimension with respect to initial data u_0 . Assume that $(u_m)_{m \in \mathbb{N}}$ is uniformly bounded with respect to m in L^∞ and u_m converges almost everywhere in $\mathbb{R} \times \mathbb{R}^+$ against some function u . Then u is the uniquely defined entropy weak solution of (1a, 1b).

With the lack of a BV estimate for the approximate solution in multiple space dimensions, one cannot expect a passage to the limit of the nonlinear terms in the entropy inequality in the classical sense, that is, the limit of u_m will not in general be a weak solution. Nevertheless, the weak compactness obtained by the L^∞ -estimate is enough to obtain a measure-valued or entropy process solution in the limit.

The key theorem for this convergence result is the following compactness theorem of Tartar; see Tartar (1983) and Eymard, Galluot and Herbin (2000).

Theorem 11 (Tartar's theorem) Let $(u_m)_{m \in \mathbb{N}}$ be a family of bounded functions in $L^\infty(\mathbb{R}^d)$. Then, there exists a subsequence $(u_{m_k})_{k \in \mathbb{N}}$ and a function $u \in L^\infty(\mathbb{R}^d \times (0, 1))$ such that for all functions $g \in C(\mathbb{R})$ the weak-* limit of $g(u_{m_k})$ exists and

$$\lim_{m \rightarrow \infty} \int_{\mathbb{R}^d} g(u_m(x)) \phi(x) \, dx = \int_0^1 \int_{\mathbb{R}^d} g(u(x, \alpha)) \phi(x) \, dx \, d\alpha, \quad \text{for all } \phi \in L^1(\mathbb{R}^d) \quad (43)$$

In order to prove the convergence of a finite volume method, it now remains to show that the residual of the entropy inequality (12) for the approximate solution u_h tends to zero if h and Δt tend to zero. Before presenting this estimate for the finite volume approximation, a general convergence theorem is given, which can be viewed as a generalization of the classical Lax-Wendroff result; see Eymard, Galluot and Herbin (2000).

Theorem 12 (Sufficient condition for convergence) Let $u_0 \in L^\infty(\mathbb{R}^d)$ and $f \in C^1(\mathbb{R})$. Further, let $(u_m)_{m \in \mathbb{N}}$ be any family of uniformly bounded functions in $L^\infty(\mathbb{R}^d \times \mathbb{R}^+)$ that satisfies the following estimate for the residual of the entropy inequality using the class of Kruzkov entropies η_κ (see Note 1).

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^+} \left(\eta_\kappa(u_m) \partial_t \phi + F_{\eta_\kappa}(u_m) \cdot \nabla \phi \right) \, dx \, dt + \int_{\mathbb{R}^d} \eta_\kappa(u_0) \phi(x, 0) \, dx \leq -R(\kappa, u_m, \phi) \quad (44)$$

for all $\kappa \in \mathbb{R}$ and $\phi \in C_0^\infty(\mathbb{R}^d \times \mathbb{R}^+, \mathbb{R}^+)$ where the residual $R(\kappa, u_m, \phi)$ tends to zero for $m \rightarrow \infty$ uniformly in κ . Then, u_m converges strongly to the unique entropy weak solution of (1a, 1b) in $L_{\text{loc}}^p(\mathbb{R}^d \times \mathbb{R}^+)$ for all $p \in [1, \infty)$.

Theorem 13 (Estimate on the residual of the entropy inequality) Let $(u_m)_{m \in \mathbb{N}}$ be a sequence of monotone finite volume approximations satisfying a local CFL-like condition as given in (39) such that $h, \Delta t$ tend to zero for $m \rightarrow \infty$. Then, there exist measures $\mu_m \in \mathcal{M}(\mathbb{R}^d \times \mathbb{R}^+)$ and $\nu_m \in \mathcal{M}(\mathbb{R}^d)$ such that the residual $R(\kappa, u_m, \phi)$ of the entropy inequality is estimated by

$$R(\kappa, u_m, \phi) \leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} (|\partial_t \phi(x, t)| + |\nabla \phi(x, t)|) \, d\mu_m(x, t) + \int_{\mathbb{R}^d} \phi(x, 0) \, d\nu_m(x)$$

for all $\kappa \in \mathbb{R}$ and $\phi \in C_0^\infty(\mathbb{R}^d \times \mathbb{R}^+, \mathbb{R}^+)$. The measures μ_m and ν_m satisfy the following properties:

1. For all compact subsets $\Omega \subset \mathbb{R}^d \times \mathbb{R}^+$, $\lim_{m \rightarrow \infty} \mu_m(\Omega) = 0$.
2. For all $g \in C_0(\mathbb{R}^d)$ the measure ν_m is given by $(\nu_m, g) = \int_{\mathbb{R}^d} g(x) |u_0(x) - u_m(x, 0)| dx$.

These theorems are sufficient for establishing convergence of monotone finite volume schemes.

Corollary 1 (Convergence theorem) Let $(u_m)_{m \in \mathbb{N}}$ be a sequence of monotone finite volume approximations satisfying the assumptions of Theorem 13. Then, u_m converges strongly to the unique entropy weak solution of (1a, 1b) in $L^\infty_{loc}(\mathbb{R}^d \times \mathbb{R}^+)$ for all $p \in [1, \infty)$.

Convergence of higher-order finite volume schemes can also be proven within the given framework as long as they are L^∞ -stable and allow for an estimate on the entropy residual in the sense of Theorem 13; for details see Kröner, Noelle and Rokyta (1995) and Chainais-Hillairet (2000).

2.2.3 Error estimates and convergence rates

There are two primary approaches taken to obtain error estimates for approximations of scalar nonlinear conservation laws. One approach is based on the ideas of Oleinik (1963) and Tadmor (1991). The second approach, which is widely used in the numerical analysis of conservation laws is based on the doubling of variables technique of Kruzkov; see Kruzkov (1970) and Kuznetsov (1976). In essence, this technique enables one to estimate the error between the exact and approximate solution of a conservation law in terms of the entropy residual $R(\kappa, u_m, \Phi)$ introduced in (44). Thus, an a posteriori error estimate is obtained. Using a priori estimates of the approximate solution (see Section 2.2.1, and Theorems 8, 9), a convergence rate or an a priori error estimate is then obtained. The next theorem gives a fundamental error estimate for conservation laws independent of the particular finite volume scheme; see Eymard, Galluët and Herbin (2000), Chainais-Hillairet (1999), and Kröner and Ohlberger (2000).

Theorem 14 (Fundamental error estimate) Let $u_0 \in BV(\mathbb{R}^d)$ and let u be an entropy weak solution of (1a, 1b). Furthermore, let $v \in L^\infty(\mathbb{R}^d \times \mathbb{R}^+)$ be a solution of the following entropy inequalities with residual term R :

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^+} \eta_\kappa(v) \partial_t \Phi + F_{\eta_\kappa}(v) \cdot \nabla \Phi + \int_{\mathbb{R}^d} \eta_\kappa(u_0) \Phi(\cdot, 0) \geq -R(\kappa, v, \Phi) \quad (45)$$

for all $\kappa \in \mathbb{R}$ and $\Phi \in C_0^1(\mathbb{R}^d \times \mathbb{R}^+; \mathbb{R}^+)$. Suppose that there exist measures $\mu_\kappa \in \mathcal{M}(\mathbb{R}^d \times \mathbb{R}^+)$ and $\nu_\kappa \in \mathcal{M}(\mathbb{R}^d)$ such

that $R(\kappa, v, \Phi)$ can be estimated independently of κ by

$$R(\kappa, v, \Phi) \leq (|\partial_t \Phi| + |\nabla \Phi|, \mu_\kappa) + (|\Phi(\cdot, 0)|, \nu_\kappa) \quad (46)$$

Let $K \subset \mathbb{R}^d \times \mathbb{R}^+$, $\omega = \text{Lip}(f)$, and choose T, R and x_0 such that $T \in]0, (R/\omega)[$ and K lies within its cone of dependence D_0 , that is, $K \subset D_0$ where D_0 is given as

$$D_0 := \bigcup_{0 \leq t \leq T} B_{R-\omega t+\delta}(x_0) \times \{t\} \quad (47)$$

Then, there exists a $\delta \geq 0$ and positive constants C_1, C_2 such that u, v satisfy the following error estimate

$$\|u - v\|_{L^1(K)} \leq T \left(\nu_\delta(B_{R+\delta}(x_0)) + C_1 \mu_\delta(D_\delta) + C_2 \sqrt{\mu_\delta(D_\delta)} \right) \quad (48)$$

This estimate can be used either as an a posteriori control of the error, as the right-hand side of the estimate (48) only depends on v , or it can be used as an a priori error bound if one is able to estimate further the measures μ_δ and ν_δ using some a priori bounds on v . Finally, note that comparable estimates to (48) are obtainable in an $L^\infty(0, T; L^1(\mathbb{R}^d))$ -norm; see Cockburn and Gau (1995) and Bouchut and Perthame (1998).

2.2.4 A posteriori error estimate

Theorem 15 (A posteriori error estimate) Assume the conditions and notations as in Theorem 14. Let $v = u_h$ be a numerical approximation to (1a, 1b) obtained from a monotone finite volume scheme that satisfies a local CFL-like condition as given in (39). Then the following error estimate holds

$$\int_K |u - u_h| \leq T (\|u_0 - u_h(\cdot, 0)\|_{L^1(B_{R+\delta}(x_0))} + C_1 \eta + C_2 \sqrt{\eta}) \quad (49)$$

where

$$\begin{aligned} \eta &= \sum_{n \in I_0} \sum_{j \in \mathcal{M}(n)} |u_j^{n+1} - u_j^n| \Delta x^n h_j^n \\ &\quad + 2 \sum_{n \in I_0} \sum_{(j,l) \in E(n)} \Delta x^n (\Delta x^n + h_{jl}) \\ &\quad \times Q_{jl}(u_j^n, u_l^n) |u_j^n - u_l^n| \end{aligned}$$

$$Q_{jl}(u, v) = \frac{2g_{jl}(u, v) - g_{jl}(u, u) - g_{jl}(v, v)}{u - v}$$

with the index sets $I_0, M(t), E(t)$ given by

$$I_0 = \{n | 0 \leq t^n \leq \min \left\{ \frac{R+\delta}{\omega}, T \right\}\}$$

$$M(t) = \{j | \text{there exists } x \in T_j \text{ such that } (x, t) \in D_{R+\delta}\}$$

$$E(t) = \{(j, l) | \text{there exists } x \in T_j \cup T_l \text{ such that } (x, t) \in D_{R+\delta}\}$$

Furthermore, the constants C_1, C_2 only depend on $T, \omega, \|u_0\|_{BV}$ and $\|u_0\|_{L^\infty}$; for details see Kröner and Ohlberger (2000).

Note that this a posteriori error estimate is local, since the error on a compact set K is estimated by discrete quantities that are supported in the cone of dependence $D_{R+\delta}$.

2.2.5 A priori error estimate

Using the weak BV estimate (Theorem 8) and the Lipschitz estimate in time (Theorem 9), the right-hand side of the a posteriori error estimate (49) can be further estimated. This yields an a priori error estimate as stated in the following theorem; for details see Cockburn and Gresham (1996), Cockburn and Gresham (1997), Cockburn, Gresham and Yang (1998), Chainais-Hillairet (1999), and Eymard, Galluët and Herbin (2000).

Theorem 16 (A priori error estimate) Assume the conditions and notations as in Theorem 14 and let $v = u_h$ be the approximation to (1a), (1b) given by a monotone finite volume scheme that satisfies a local CFL-like condition as given in (39). Then there exists a constant $C \geq 0$ such that

$$\int_K |u - u_h| dx \leq C h^{1/4}$$

Moreover, in the one-dimensional case, the optimal convergence rate of $h^{1/2}$ is obtained.

2.2.6 Convergence proofs via the streamline diffusion discontinuous Galerkin finite element method

It is straightforward to show that the fully discrete finite volume scheme (22) can be viewed as a specific case of the more general streamline diffusion discontinuous Galerkin (SD-DG) finite element method which utilizes the mesh dependent broken space V_h^s defined as

$$V_h^s = \{v | v|_T \in \mathcal{P}_p(T), \quad \forall T \in \mathcal{T}\} \quad (50)$$

with $\mathcal{P}_p(T)$ the space of polynomials of degree $\leq p$ in the control volume T . By generalizing the notion of gradient and flux to include the time coordinate as well, the discontinuous Galerkin finite element method for a space-time tessellation \mathcal{T}^s spanning the time slab increment $[t^n, t^{n+1}]$ is given compactly by the following variational statement.

SD-DG(p) finite element method. Find $u_h \in V_h^s$ such that $\forall v_h \in V_h^s$ and $n = 0, 1, \dots$

$$\begin{aligned} \sum_{T \in \mathcal{T}^s} \left(\int_T (v_h + \delta(u_h)) f'(u_h) \cdot \nabla u_h \nabla \cdot f(u_h) dx \right. \\ \left. + \int_T \tilde{e}(u_h) \nabla u_h \cdot \nabla v_h dx \right. \\ \left. + \int_{\partial T} v_h (g(\tilde{v}; u_{-h}, u_{+h}) - f(u_{-h}) \cdot \tilde{v}) ds \right) = 0 \end{aligned} \quad (51)$$

where \tilde{v} denotes the unit exterior normal on ∂T . In the integration over ∂T , it is understood that for the portion $x \in \partial T \cap \partial T'$ that u_{-h}, v_{-h} denotes the trace restriction of $u_h(T)$ and $v_h(T)$ onto ∂T and u_{+h} denotes the trace restriction of $u_h(T')$ onto $\partial T'$. Given this space-time formulation, convergence results for a scalar nonlinear conservation law in multidimensions and unstructured meshes are given in Jaffre, Johnson and Szepessy (1995) for specific choices of the stabilization functions $\delta(u_h): \mathbb{R} \rightarrow \mathbb{R}^+$ and $\tilde{e}(u_h): \mathbb{R} \rightarrow \mathbb{R}^+$ together with a monotone numerical flux function $g(\tilde{v}; u_{-h}, u_{+h})$. Using their stabilization functions together with a monotone flux function, the following convergence result is obtained:

Theorem 17 (SD-DG(p) convergence) Suppose that components of $f'(u) \in C^0(\mathbb{R})$ are bounded and that $u_0 \in L_2(\mathbb{R}^d)$ has compact support. Then the solution u_h of the SD-DG(p) method converges strongly in $L^\infty_p(\mathbb{R}^d \times \mathbb{R}^+)$, $1 \leq p \leq 2$, to the unique solution u of the scalar nonlinear conservation law system (1a, 1b) as $H \equiv \max(\|h\|_{L^\infty(\mathbb{R}^d)}, \Delta t)$ tends to zero.

The proof of convergence to a unique entropy solution on general meshes for $p \geq 0$ is based on an extension by Szepessy (1989) of a uniqueness result by DiPerna (1985) by providing convergence for a sequence of approximations satisfying:

- a uniform L_∞ bound in time and L_2 in space,
- entropy consistency and inequality for all Kruzkov entropies,
- consistency with initial data.

By choosing SD-DG(0), the dependence on the as yet unspecified stabilization functions $\delta(u_h)$ and $\tilde{e}(u_h)$ vanishes

identically and the fully discrete scheme (22) with monotone flux function is *exactly* reproduced, thus yielding a convergence proof for general scalar conservation laws for the finite volume method as well. Subsequently, Cockburn and Gremaud (1996) replaced the $L^\infty(L^2)$ norm analysis of Jaffre, Johnson and Szepessy (1995) with an alternate analysis in $L^\infty(L^1)$ thus yielding $O(h^{1/8})$ and $O(h^{1/4})$ error estimates in time and space respectively.

3 HIGHER-ORDER ACCURATE FV GENERALIZATIONS

Even for linear advection problems where an $O(h^{1/2})$ L_2 -norm error bound for the monotone flux schemes of Section 2 is known to be sharp (Peterson, 1991), an $O(h)$ solution error is routinely observed in numerical experiments on smoothly varying meshes with convex flux functions. Nevertheless, first-order accurate schemes are generally considered too inaccurate for most quantitative calculations unless the mesh spacing is made excessively small, thus rendering the schemes inefficient. Godunov (1959) has shown that all linear schemes that preserve solution monotonicity are at most first-order accurate. The low-order accuracy of these monotonicity preserving linear schemes has motivated the development of higher-order accurate schemes with the important distinction that these new schemes utilize essential *nonlinearity* so that monotone resolution of discontinuities and high-order accuracy away from discontinuities are simultaneously attained.

3.1 Higher-order accurate FV schemes in 1-D

A significant step forward in the generalization of Godunov's finite volume method to higher-order accuracy is due to van Leer (1979). In the context of Lagrangian hydrodynamics with Eulerian remapping, van Leer generalized

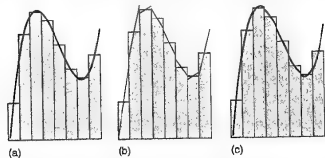


Figure 3. Piecewise polynomial approximation used in the finite volume method (a) cell averaging of analytic data, (b) piecewise linear reconstruction from cell averages and (c) piecewise quadratic reconstruction from cell averages.

Godunov's method by employing linear solution reconstruction in each cell (see Figure 3b). Let N denote the number of control volume cells in space so that the j th cell extends over the interval $T_j = [x_{j-1/2}, x_{j+1/2}]$ with length Δx_j such that $\cup_{j=1}^N T_j = [0, 1]$ with $T_i \cap T_j = \emptyset$, $i \neq j$. In a purely Eulerian setting, the higher-order accurate schemes of van Leer are of the form

$$\frac{du}{dt} + \frac{1}{\Delta x_j} (g(u_{j+1/2}^*, u_{j+1/2}^*) - g(u_{j-1/2}^*, u_{j-1/2}^*)) = 0$$

where $g(u, v)$ is a numerical flux function utilizing states $u_{j+1/2}^*$ and $u_{j-1/2}^*$ obtained from evaluation of the linear solution reconstructions from the left and right cells surrounding the interfaces $x_{j+1/2}$. By altering the slope of the linear reconstruction in cells, nonoscillatory resolution of discontinuities can be obtained. Note that although obtaining the exact solution of the scalar nonlinear conservation law with linear initial data is a formidable task, the solution at each cell interface location for small enough time is the same as the solution of the Riemann problem with piecewise constant data equal to the linear solution approximation evaluated at the same interface location. Consequently, the numerical flux functions used in Section 2 can be once again used in the generalized schemes of van Leer. This single observation greatly simplifies the construction of higher-order accurate generalizations of Godunov's method. This observation also suggested a relatively straightforward extension of van Leer ideas to quadratic approximation in each cell (see Figure 3c) as discussed in early work by Colella and Woodward (1984). Although these generalizations of Godunov's method and further generalizations given later can be interpreted in 1-D as finite difference discretizations, concepts originally developed in 1-D, such as solution monotonicity, positive coefficient discretization, and discrete maximum principle analysis are often used in the design of finite volume methods in multiple space dimensions and on unstructured meshes where finite difference discretization is problematic.

3.1.1 TVD schemes

In considering the scalar nonlinear conservation law (1a, 1b), Lax (1973) made the following basic observation:

the total increasing and decreasing variations of a differentiable solution between any pair of characteristics are conserved

Furthermore, in the presence of shock wave discontinuities, information is lost and the total variation decreases. For the 1-D nonlinear conservation law with compactly supported (or periodic) solution data $u(x, t)$, integrating along the

constant time spatial coordinate at times t_1 and t_2 yields

$$\int_{-\infty}^{\infty} |du(x, t_2)| \leq \int_{-\infty}^{\infty} |du(x, t_1)|, \quad t_2 \geq t_1 \quad (52)$$

This motivated Harten (1983) to consider the discrete total variation

$$TV(u_h) \equiv \sum_j |\Delta_{j+1/2} u_h|, \quad \Delta_{j+1/2} u_h \equiv u_{j+1} - u_j$$

and the discrete total variation nonincreasing (TVNI) bound counterpart to (52)

$$TV(u_h^{n+1}) \leq TV(u_h^n) \quad (53)$$

in the design of numerical discretizations for nonlinear conservation laws. A number of simple results relating TVNI schemes and monotone schemes follow from simple analysis.

Theorem 18 (TVNI and monotone scheme properties, Harten, 1983) (i) Monotone schemes are TVNI, (ii) TVNI schemes are monotonicity preserving, that is, the number of solution extrema is preserved in time.

Property (i) follows from the L_1 -contraction property of monotone schemes. Property (ii) is readily shown using a proof by contradiction, by assuming a TVNI scheme with monotone initial data that produces new solution data at a later time with interior solution extrema present. Using the notion of discrete total variation, Harten (1983) then constructed sufficient algebraic conditions for achieving the TVNI inequality (53).

Theorem 19 (Harten's explicit TVD criteria) The fully discrete explicit 1-D scheme

$$u_j^{n+1} = u_j^n + \Delta t \left(C_{j+1/2} (u_h^n) \Delta_{j+1/2} u_h^n + D_{j+1/2} (u_h^n) \Delta_{j-1/2} u_h^n \right), \quad j = 1, \dots, N \quad (54)$$

is total variation nonincreasing if for each j

$$C_{j+1/2} \geq 0 \quad (55a)$$

$$D_{j+1/2} \leq 0 \quad (55b)$$

$$1 - \Delta t (C_{j-1/2} - D_{j+1/2}) \geq 0 \quad (55c)$$

Note that although the inequality constraints (55a–55c) in Theorem 19 insure that the total variation is nonincreasing, these conditions are often referred to as total variation diminishing (TVD) conditions. Also note that inequality (55c) implies a CFL-like time step restriction that may be

more restrictive than the time step required for stability of the numerical method. The TVD conditions are easily generalized to wider support stencils written in incremental form; see, for example, Jameson and Lax (1986) and their corrected result in Jameson and Lax (1987).

While this simple Euler explicit time integration scheme may seem too crude for applications requiring true high-order space-time accuracy, special attention and analysis is given to this fully discrete form because it serves as a fundamental building block for an important class of high-order accurate Runge–Kutta time integration techniques discussed in Section 4.1 that, by construction, inherit TVD (and later maximum principle) properties of the fully discrete scheme (54).

Theorem 20 (Generalized explicit TVD criteria) The fully discrete explicit 1-D scheme

$$u_j^{n+1} = u_j^n + \Delta t \sum_{l=k}^{k-1} C_{j+1/2}^{(l)} (u_h^n) \Delta_{j+1/2} u_h^n, \quad j = 1, \dots, N \quad (56)$$

with integer stencil width parameter $k > 0$ is total variation nonincreasing if for each j

$$C_{j+1/2}^{(k-1)} \geq 0 \quad (57a)$$

$$C_{j+1/2}^{(k)} \leq 0 \quad (57b)$$

$$C_{j+1/2}^{(l-1)} - C_{j+1/2}^{(l)} \geq 0, \quad -k+1 < l < k-1, \quad l \neq 0 \quad (57c)$$

$$1 - \Delta t (C_{j-1/2}^{(0)} - C_{j+1/2}^{(k-1)}) \geq 0 \quad (57d)$$

The extension to implicit methods follows immediately upon rewriting the implicit scheme in terms of the solution spatial increments $\Delta_{j+1/2} u_h$ and imposing sufficient algebraic conditions such that the implicit matrix acting on spatial increments has a nonnegative inverse.

Theorem 21 (Generalized implicit TVD criteria) The fully discrete implicit 1-D scheme

$$u_j^{n+1} - \Delta t \sum_{l=k}^{k-1} C_{j+1/2}^{(l)} (u_h^{n+1}) \Delta_{j+1/2} u_h^{n+1} = u_j^n, \quad j = 1, \dots, N \quad (58)$$

with integer stencil width parameter $k > 0$ is total variation nonincreasing if for each j

$$C_{j+1/2}^{(k-1)} \geq 0 \quad (59a)$$

$$C_{j+1/2}^{(k)} \leq 0 \quad (59b)$$

$$C_{j+1/2}^{(l-1)} - C_{j+1/2}^{(l)} \geq 0, \quad -k+1 \leq l \leq k-1, \quad l \neq 0 \quad (59c)$$

Theorems 20 and 21 provide sufficient conditions for non-increasing total variation of explicit (56) or implicit (58) numerical schemes written in incremental form. These incremental forms do not imply *discrete conservation* unless additional constraints are imposed on the discretizations. A sufficient condition for discrete conservation of the discretizations (56) and (58) is that these discretizations can be written in a finite volume flux balance form

$$g_{j+1/2} - g_{j-1/2} = \sum_{l=k}^{k+1} C_{j+1/2}^{(l)}(u_h) \Delta_{j+1/2} u_h$$

where $g_{j+1/2}$ are the usual numerical flux functions. Section 3.1.2 provides an example of how the discrete TVD conditions and discrete conservation can be simultaneously achieved. A more comprehensive overview of finite volume numerical methods based on TVD constructions can be found in the books by Godlewski and Raviart (1991) and LeVeque (2002).

3.1.2 MUSCL schemes

A general family of TVD discretizations with 5-point stencil is the monotone upstream-centered scheme for conservation laws (MUSCL) discretization of van Leer (1979) and van Leer (1985). MUSCL schemes utilize a κ -parameter family of interpolation formulas with *limiter function* $\Psi(R): \mathbb{R} \rightarrow \mathbb{R}$

$$\begin{aligned} u_{j+1/2}^- &= u_j + \frac{1+\kappa}{4} \Psi(R_j) \Delta_{j-1/2} u_h \\ &\quad + \frac{1-\kappa}{4} \Psi\left(\frac{1}{R_j}\right) \Delta_{j+1/2} u_h \\ u_{j-1/2}^+ &= u_j - \frac{1+\kappa}{4} \Psi\left(\frac{1}{R_j}\right) \Delta_{j-1/2} u_h \\ &\quad - \frac{1-\kappa}{4} \Psi(R_j) \Delta_{j+1/2} u_h \end{aligned} \quad (60)$$

where R_j is a ratio of successive solution increments

$$R_j = \frac{\Delta_{j+1/2} u_h}{\Delta_{j-1/2} u_h} \quad (61)$$

The technique of incorporating limiter functions to obtain nonoscillatory resolution of discontinuities and steep gradients dates back to Boris and Book (1973). For convenience, the interpolation formulas (60) have been written for a uniformly spaced mesh, although the extension to irregular mesh spacing is straightforward. The unlimited form of this interpolation is obtained by setting $\Psi(R) = 1$. In this

unlimited case, the truncation error for the conservation law divergence in (1a) is given by

$$\text{Truncation Error} = -\frac{[\kappa - (1/3)]}{4} (\Delta x)^2 \frac{\partial^3}{\partial x^3} f(u)$$

This equation reveals that for $\kappa = 1/3$, the 1-D MUSCL formula yields an overall spatial discretization with $O(\Delta x^3)$ truncation error. Using the MUSCL interpolation formulas given in (60), sufficient conditions for the discrete TVD property are easily obtained.

Theorem 22 (MUSCL TVD scheme) *The fully discrete 1-D scheme*

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x_j} (g_{j+1/2}^n - g_{j-1/2}^n), \quad j = 1, \dots, N$$

with monotone Lipschitz continuous numerical flux function

$$g_{j+1/2} = g(u_{j+1/2}^-, u_{j+1/2}^+)$$

utilizing the κ -parameter family of MUSCL interpolation formulas (60) and (61) is total variation nonincreasing if there exists a $\Psi(R)$ such that $\forall R \in \mathbb{R}$

$$0 \leq \Psi(R) \leq \frac{3-\kappa}{1-\kappa} - (1+\alpha) \frac{1+\kappa}{1-\kappa} \quad (62a)$$

and

$$0 \leq \frac{\Psi(R)}{R} \leq 2 + \alpha \quad (62b)$$

with $\alpha \in [-2, 2(1-\kappa)/(1+\kappa)]$ under the time step restriction

$$1 - \frac{\Delta t}{\Delta x_j} \frac{2 - (2+\alpha)\kappa}{1-\kappa} \left| \frac{\partial g}{\partial u} \right|_{j+1/2}^{\max} \geq 0$$

where

$$\begin{aligned} \left| \frac{\partial g}{\partial u} \right|_{j+1/2}^{\max} &= \sup_{\substack{\tilde{u} \in [u_{j-1/2}^-, u_{j+1/2}^+] \\ \tilde{v} \in [u_{j-1/2}^+, u_{j+1/2}^-]}} \left(\frac{\partial g}{\partial u^-}(\tilde{u}, u_{j+1/2}^+) \right. \\ &\quad \left. - \frac{\partial g}{\partial u^+}(u_{j-1/2}^-, \tilde{v}) \right) \end{aligned}$$

For accuracy considerations away from extrema, it is desirable that the unlimited form of the discretization is obtained. Consequently, the constraint $\Psi(1) = 1$ is also imposed upon the limiter function. This constraint together with the algebraic conditions (62a, b) are readily achieved using the well-known *MinMod* limiter, Ψ^{MM} , with compression parameter β determined from the TVD

Table 1. Members of the MUSCL TVD family of schemes.

| κ | Unlimited scheme | β_{\max} | Truncation error |
|----------|----------------------|----------------|--|
| 1/3 | Third-order | 4 | 0 |
| -1 | Fully upwind | 2 | $\frac{1}{3}(\Delta x)^2 \frac{\partial^3}{\partial x^3} f(u)$ |
| 0 | Fromm's | 3 | $\frac{1}{12}(\Delta x)^2 \frac{\partial^3}{\partial x^3} f(u)$ |
| 1/2 | Low truncation error | 5 | $-\frac{1}{24}(\Delta x)^2 \frac{\partial^3}{\partial x^3} f(u)$ |

analysis

$$\Psi^{\text{MM}}(R) = \max(0, \min(R, \beta)), \quad \beta \in \left[1, \frac{(3-\kappa)}{(1-\kappa)}\right]$$

Table 1 summarizes the MUSCL scheme and maximum compression parameter for a number of familiar discretizations. Another limiter due to van Leer that meets the technical conditions of Theorem 22 and also satisfies $\Psi(1) = 1$ is given by

$$\Psi^{\text{VL}}(R) = \frac{R + |R|}{1 + |R|}$$

This limiter exhibits differentiability away from $R = 0$, which improves the iterative convergence to steady state for many algorithms. Numerous other limiter functions are considered and analyzed in Sweby (1984).

Unfortunately, TVD schemes locally degenerate to piecewise constant approximations at smooth extrema, which locally degrades the accuracy. This is an unavoidable consequence of the strict TVD condition.

Theorem 23 (TVD critical point accuracy, Osher, 1984) *The TVD discretizations (54), (56) and (58) all reduce to at most first-order accuracy at nonsonic critical points, that is, points u^* at which $f'(u^*) \neq 0$ and $u_x^* = 0$.*

3.1.3 ENO/WENO schemes

To circumvent the degradation in accuracy of TVD schemes at critical points, weaker constraints on the solution total variation were devised. To this end, Harten proposed the following abstract framework for generalized Godunov schemes in operator composition form (see Harten *et al.*, 1986, 1987; Harten, 1989)

$$u_h^{n+1} = A \cdot E(\tau) \cdot R_p^0(\cdot; u_h^n) \quad (63)$$

In this equation, $u_h^n \in V_h^0$ denotes the global space of piecewise constant cell averages as defined in (20), $R_p^0(x)$ is a reconstruction operator, which produces a cell-wise

discontinuous p th order polynomial approximation from the given solution cell averages, $E(\tau)$ is the evolution operator for the PDE (including boundary conditions), and A is the cell averaging operator. Since A is a nonnegative operator and $E(\tau)$ represents exact evolution in the small, the control of solution oscillations and Gibbs-like phenomena is linked directly to oscillation properties of the reconstruction operator, $R_p^0(x)$. One has formally in one space dimension

$$\text{TV}(u_h^{n+1}) = \text{TV}(A \cdot E(\tau) \cdot R_p^0(\cdot; u_h^n)) \leq \text{TV}(R_p^0(x; u_h^n))$$

so that the total variation depends entirely upon properties of the reconstruction operator $R_p^0(x; u_h^n)$. The requirements of high-order accuracy for smooth solutions and discrete conservation give rise to the following additional design criterion for the reconstruction operator

$$\bullet R_p^0(x; u_h) = u(x) + e(x) \Delta x^{p+1} + O(\Delta x^{p+2}) \quad \text{whenever } u \text{ is smooth} \quad (64a)$$

$$\bullet A|_{\mathcal{T}_h} R_p^0(x; u_h) = u_h|_{\mathcal{T}_h} = u_j, \quad j = 1, \dots, N \quad \text{to insure discrete conservation} \quad (64b)$$

$$\bullet \text{TV}(R(x; u_h^n)) \leq \text{TV}(u_h^n) + O(\Delta x^{p+1}) \quad \text{an essentially nonoscillatory reconstruction.} \quad (64c)$$

Note that $e(x)$ may not be Lipschitz continuous at certain points so that the cumulative error in the scheme is $O(\Delta x^p)$ in a maximum norm but remains $O(\Delta x^{p+1})$ in an L_1 -norm. To achieve the requirements of (64a–64c), Harten and coworkers considered breaking the task into two parts

- Polynomial reconstruction from a given stencil of cell averages
- Construction of a “smoothest” polynomial approximation by a solution adaptive stencil selection algorithm.

In the next section, a commonly used reconstruction technique from cell averages is considered. This is then followed by a description of the solution adaptive stencil algorithm proposed by Harten *et al.* (1986).

3.1.4 Reconstruction via primitive function

Given cell averages u_j of a piecewise smooth function $u(x)$, one can inexpensively evaluate pointwise values of the *primitive function* $U(x)$

$$U(x) = \int_{x_0}^x u(\xi) d\xi$$

by exploiting the relationship

$$\sum_{j=j_0}^j \Delta x_j u_j = U(x_{j+1/2})$$

Let $H_p(x; u)$ denote a p th order piecewise polynomial interpolant of a function u . Since

$$u(x) = \frac{d}{dx} U(x)$$

an interpolant of the primitive function given pointwise samples $U(x_{j+1/2})$ yields a reconstruction operator

$$R_p^0(x; u_k) = \frac{d}{dx} H_{p+1}(x; U)$$

As a polynomial approximation problem, whenever $U(x)$ is smooth one obtains

$$\frac{d^k}{dx^k} H_p(x; U) = \frac{d^k}{dx^k} U(x) + O(\Delta x^{p+1-k}), 0 \leq k \leq p$$

and consequently

$$\frac{d^l}{dx^l} R_p^0(x; u_k) = \frac{d^l}{dx^l} u(x) + O(\Delta x^{p+1-l})$$

By virtue of the use of the primitive function $U(x)$, it follows that

$$A|_{\tau} R_p^0(x; u_k) = u_j$$

and from the polynomial interpolation problem for smooth data

$$R_p^0(x; u_k) = u(x) + O(\Delta x^{p+1})$$

as desired.

3.1.5 ENO reconstruction

The reconstruction technique outlined in Section 3.1.4 does not satisfy the oscillation requirement given in (64c). This motivated Harten and coworkers to consider a new algorithm for essentially nonoscillatory (ENO) piecewise polynomial interpolation. When combined with the reconstruction technique of Section 3.1.4, the resulting reconstruction technique satisfies (64a–c). Specifically, a new interpolant $H_p(x; u)$ is constructed so that when applied to piecewise smooth data $u(x)$ gives high-order accuracy

$$\frac{d^k}{dx^k} H_p(x; v) = \frac{d^k}{dx^k} v(x) + O(\Delta x^{p+1-k}), 0 \leq k \leq p$$

but avoids having Gibbs oscillations at discontinuities in the sense

$$TV(H_p(x; v)) \leq TV(v) + O(\Delta x^{p+1})$$

The strategy pursued by Harten and coworkers was to construct such an ENO polynomial $H_p(x; w)$ using the following steps. Define

$$H_p^{\text{ENO}}(x; w) = P_{p,j+1/2}^{\text{ENO}}(x; w) \quad \text{for } x_j \leq x \leq x_{j+1}, \\ j = 1, \dots, N$$

where $P_{p,j+1/2}^{\text{ENO}}$ is the p th degree polynomial which interpolates $w(x)$ at the $p+1$ successive points $\{x_i\}$, $i_p(j) \leq i \leq i_p(j) + p$ that include x_j and x_{j+1} , that is,

$$P_{p,j+1/2}^{\text{ENO}}(x_i; w) = w(x_i), \quad i_p(j) \leq i \leq i_p(j) + p, \\ 1 - p \leq i_p(j) - j \leq 0 \quad (65)$$

Equation (65) describes p possible polynomials depending on the choice of $i_p(j)$ for an interval (x_j, x_{j+1}) . The ENO strategy selects the value $i_p(j)$ for each interval that produces the 'smoothest' polynomial interpolant for a given input data. More precisely, information about smoothness of $w(x)$ is extracted from a table of divided differences of $w(x)$ defined recursively for $i = 1, \dots, N$ by

$$w[x_i] = w(x_i) \\ w[x_i, x_{i+1}] = \frac{w[x_{i+1}] - w[x_i]}{x_{i+1} - x_i} \\ \vdots \\ w[x_i, \dots, x_{i+k}] = \frac{w[x_{i+1}, \dots, x_{i+k}] - w[x_i, \dots, x_{i+k-1}]}{x_{i+k} - x_i}$$

The stencil producing the smoothest interpolant is then chosen hierarchically by setting

$$i_1(j) = j$$

and for $1 \leq k \leq p-1$

$$i_{k+1}(j) = \begin{cases} i_k(j) - 1 & \text{if } |w[x_{i_k(j)-1}, \dots, w[x_{i_k(j)+k}]]| \\ < |w[x_{i_k(j)}, \dots, w[x_{i_k(j)+k+1}]]| \\ i_k(j) & \text{otherwise} \end{cases} \quad (66)$$

Harten *et al.* (1986) demonstrate the following properties of this ENO interpolation strategy

- The accuracy condition

$$P_{p,j+1/2}^{\text{ENO}}(x) = w(x) + O(\Delta x^{p+1}), \quad x \in (x_j, x_{j+1})$$

- $P_p^{\text{ENO}}(x)$ is monotone in any cell interval containing a discontinuity.

- There exists a function $z(x)$ nearby $P_p^{\text{ENO}}(x)$ in the interval (x_j, x_{j+1}) in the sense

$$z(x) = P_{p,j+1/2}^{\text{ENO}}(x) + O(\Delta x^{p+1}), \quad x \in (x_j, x_{j+1})$$

that is total variation bounded, that is, the nearby function $z(x)$ satisfies

$$TV(z) \leq TV(w)$$

3.1.6 WENO reconstruction

The solution adaptive nature of the ENO stencil selection algorithm (66) yields nondifferentiable fluxes that impede convergence to steady state. In addition, the stencil selection algorithm chooses only one of p possible stencils and other slightly less smooth stencils may give similar accuracy. When $w(x)$ is smooth, using a linear combination of all p stencils with optimized weights yields a more accurate $O(\Delta x^{2p-1})$ interpolant. More specifically, let $P_{p,j+1/2}^{(k)}$ denote the unique polynomial interpolating $p+1$ points with stencil $\{x_{j+1-p+k}, x_{j+1+k}\}$ then

$$P_{p,j+1/2}(w(x)) = \sum_{k=0}^{p-1} \omega_k P_{p,j+1/2}^{(k)}(x) + O(\Delta x^{2p-1}), \\ \sum_{k=0}^{p-1} \omega_k = 1$$

For example, optimized weights for $p = 1, 2, 3$ yielding $O(\Delta x^{2p-1})$ accuracy are readily computed

$$p = 1: \quad \omega_0 = 1 \\ p = 2: \quad \omega_0 = \frac{2}{3}, \omega_1 = \frac{1}{3} \\ p = 3: \quad \omega_0 = \frac{3}{10}, \omega_1 = \frac{3}{5}, \omega_2 = \frac{1}{10}$$

In the weighted essentially nonoscillatory (WENO) schemes of Jiang and Shu (1996) and Shu (1999), approximate weights, $\tilde{\omega}_k$, are devised such that for smooth solutions

$$\tilde{\omega}_k = \omega_k + O(\Delta x^{p-1})$$

so that the $O(\Delta x^{2p-1})$ accuracy is still retained using these approximations

$$P_{p,j+1/2}(w(x)) = \sum_{k=0}^{p-1} \tilde{\omega}_k P_{p,j+1/2}^{(k)}(x) + O(\Delta x^{2p-1}), \\ \sum_{k=0}^{p-1} \tilde{\omega}_k = 1$$

The approximate weights are constructed using the ad hoc formulas

$$\alpha_k = \frac{\omega_k}{(\epsilon + \beta_k)^2}, \quad \tilde{\omega}_k = \frac{\alpha_k}{\sum_{k=0}^{p-1} \alpha_k}$$

where ϵ is an approximation to the square root of the machine precision and β_k is a smoothness indicator

$$\beta_k = \sum_{l=1}^{k-1} \int_{x_{j-1/2}}^{x_{j+1/2}} \Delta x_j^{2l-1} \left(\frac{d^l P_k^k(x)}{dx^l} \right)^2 dx$$

For a sequence of smooth solutions with decreasing smoothness indicator β_k , these formulas approach the optimized weights, $\tilde{\omega}_k \rightarrow \omega_k$. These formulas also yield vanishing weights $\tilde{\omega}_k \rightarrow 0$ for stencils with large values of the smoothness indicator such as those encountered at discontinuities. In this way, the WENO construction retains some of the attributes of the original ENO formulation but with increased accuracy in smooth solution regions and improved differentiability often yielding superior robustness for steady state calculations.

3.2 Higher-order accurate FV schemes in multidimensions

Although the one-dimensional TVD operators may be readily applied in multidimensions on a dimension-by-dimension basis, a result of Goodman and LeVeque (1985) shows that TVD schemes in two or more space dimensions are only first-order accurate.

Theorem 24 (Accuracy of TVD schemes in multidimensions) Any two-dimensional finite volume scheme of the form

$$u_{i,j}^{n+1} = u_{i,j}^n - \frac{\Delta t}{|T_{i,j}|} (g_{i+1/2,j}^n - g_{i-1/2,j}^n) \\ - \frac{\Delta t}{|T_{i,j}|} (h_{i,j+1/2}^n - h_{i,j-1/2}^n), \quad 1 \leq i \leq M, 1 \leq j \leq N$$

with Lipschitz continuous numerical fluxes for integers p, q, r, s

$$g_{i+1/2,j} = g(u_{i-p,j-q}, \dots, u_{i+r,j+s}) \\ h_{i,j+1/2} = h(u_{i-p,j-q}, \dots, u_{i+r,j+s})$$

that is total variation nonincreasing in the sense

$$TV(u_{i,j}^{n+1}) \leq TV(u_{i,j}^n)$$

where

$$TV(u) \equiv \sum_{i,j} \left[\Delta x_{i+1/2,j} |u_{i+1,j} - u_{i,j}| + \Delta x_{i,j+1/2} |u_{i,j+1} - u_{i,j}| \right]$$

is at most first-order accurate.

Motivated by the negative results of Goodman and LeVeque, weaker conditions yielding solution monotonicity preservation have been developed from discrete maximum principle analysis. These alternative constructions have the positive attribute that they extend to unstructured meshes as well.

3.2.1 Positive coefficient schemes on structured meshes

Theorem 5 considers schemes of the form

$$u_j^{n+1} = u_j^n + \frac{\Delta t}{|T_j|} \sum_{v \in \partial T_j} C_{jk}(u_k^n)(u_k^n - u_j^n), \quad \forall T_j \in \mathcal{T}$$

and provides a local space-time discrete maximum principle

$$\min_{v \in \partial T_j} (u_k^n, u_j^n) \leq u_j^{n+1} \leq \max_{v \in \partial T_j} (u_k^n, u_j^n)$$

$\forall T_j \in \mathcal{T}$ under a CFL-like condition on the time step parameter if all coefficients C_{jk} are nonnegative. Schemes of this type are often called *positive coefficient schemes* or more simply *positive schemes*. To circumvent the negative result of Theorem 24, Spekreijse (1987) developed a family of high-order accurate positive coefficient schemes on two-dimensional structured $M \times N$ meshes. For purposes of positivity analysis, these schemes are written in incremental form on a $M \times N$ logically rectangular 2-D mesh

$$\begin{aligned} u_{i,j}^{n+1} = & u_{i,j}^n + \Delta t \left(A_{i+1,j}^n (u_{i+1,j}^n - u_{i,j}^n) \right. \\ & + B_{i,j+1}^n (u_{i,j+1}^n - u_{i,j}^n) + C_{i-1,j}^n (u_{i-1,j}^n - u_{i,j}^n) \\ & \left. + D_{i,j-1}^n (u_{i,j-1}^n - u_{i,j}^n) \right), \quad 1 \leq i \leq M, 1 \leq j \leq N \end{aligned} \quad (67)$$

where the coefficients are nonlinear functions of the solution

$$\begin{aligned} A_{i+1,j}^n &= A(\dots, u_{i-1,j}^n, u_{i,j}^n, u_{i+1,j}^n, \dots) \\ B_{i,j+1}^n &= B(\dots, u_{i,j-1}^n, u_{i,j}^n, u_{i,j+1}^n, \dots) \\ C_{i-1,j}^n &= C(\dots, u_{i-1,j}^n, u_{i,j}^n, u_{i+1,j}^n, \dots) \\ D_{i,j-1}^n &= D(\dots, u_{i,j-1}^n, u_{i,j}^n, u_{i,j+1}^n, \dots) \end{aligned}$$

Once written in incremental form, the following theorem follows from standard positive coefficient maximum principle analysis.

Theorem 25 (Positive coefficient schemes in multidimensions) The discretization (67) is a positive coefficient scheme if for each $1 \leq i \leq M$, $1 \leq j \leq N$ and time slab increment $[t^n, t^{n+1}]$

$$A_{i+1,j}^n \geq 0, B_{i,j+1}^n \geq 0, C_{i-1,j}^n \geq 0, D_{i,j-1}^n \geq 0 \quad (68)$$

and

$$1 - \Delta t (A_{i+1,j}^n + B_{i,j+1}^n + C_{i-1,j}^n + D_{i,j-1}^n) \geq 0 \quad (69)$$

with discrete space-time maximum principle

$$\begin{aligned} \min(u_{i,j}^n, u_{i-1,j}^n, u_{i,j-1}^n, u_{i+1,j}^n, u_{i,j+1}^n) &\leq u_{i,j}^{n+1} \\ &\leq \max(u_{i,j}^n, u_{i-1,j}^n, u_{i,j-1}^n, u_{i+1,j}^n, u_{i,j+1}^n) \end{aligned}$$

and discrete maximum principle at steady state

$$\begin{aligned} \min(u_{i,j}^*, u_{i-1,j}^*, u_{i,j-1}^*, u_{i+1,j}^*, u_{i,j+1}^*) &\leq \\ u_{i,j}^* &\leq \max(u_{i,j}^*, u_{i-1,j}^*, u_{i,j-1}^*, u_{i+1,j}^*, u_{i,j+1}^*) \end{aligned}$$

where u^* denotes the numerical steady state.

Using a procedure similar to that used in the development of MUSCL TVD schemes in 1-D, Spekreijse (1987) developed a family of monotonicity preserving MUSCL approximations in multidimensions from the positivity conditions of Theorem 25.

Theorem 26 (MUSCL positive coefficient scheme) Assume a fully discrete 2-D finite volume scheme

$$\begin{aligned} u_{i,j}^{n+1} = & u_{i,j}^n - \frac{\Delta t}{|T_{i,j}|} (g_{i+1/2,j}^n - g_{i-1/2,j}^n) \\ & - \frac{\Delta t}{|T_{i,j}|} (h_{i,j+1/2}^n - h_{i,j-1/2}^n), \quad 1 \leq i \leq M, 1 \leq j \leq N \end{aligned}$$

utilizing monotone Lipschitz continuous numerical flux functions

$$\begin{aligned} g_{i+1/2,j} &= g(u_{i-1/2,j}^*, u_{i+1/2,j}^*) \\ h_{i,j+1/2} &= h(u_{i,j-1/2}^*, u_{i,j+1/2}^*) \end{aligned}$$

and MUSCL extrapolation formulas

$$u_{i+1/2,j}^* = u_{i,j} + \frac{1}{2} \Psi(R_{i,j})(u_{i,j} - u_{i-1,j})$$

$$u_{i-1/2,j}^* = u_{i,j} - \frac{1}{2} \Psi\left(\frac{1}{R_{i,j}}\right)(u_{i+1,j} - u_{i,j})$$

$$u_{i,j+1/2}^* = u_{i,j} + \frac{1}{2} \Psi(S_{i,j})(u_{i,j} - u_{i,j-1})$$

$$u_{i,j-1/2}^* = u_{i,j} - \frac{1}{2} \Psi\left(\frac{1}{S_{i,j}}\right)(u_{i,j+1} - u_{i,j})$$

with

$$R_{i,j} \equiv \frac{u_{i+1,j} - u_{i,j}}{u_{i,j} - u_{i-1,j}}, \quad S_{i,j} \equiv \frac{u_{i,j+1} - u_{i,j}}{u_{i,j} - u_{i,j-1}}$$

This scheme satisfies the local maximum principle properties of Lemma 25 and is second-order accurate if the limiter $\Psi = \Psi(R)$ has the properties that there exist constants $\beta \in (0, \infty)$, $\alpha \in [-2, 0]$ such that $\forall R \in \mathbb{R}$

$$\alpha \leq \Psi(R) \leq \beta, \quad -\beta \leq \frac{\Psi(R)}{R} \leq 2 + \alpha \quad (70)$$

with the constraint $\Psi(1) = 1$ and the smoothness condition $\Psi(R) \in C^2$ near $R = 1$ together with a time step restriction for stability

$$1 - (1 + \beta) \frac{\Delta t}{|T_{i,j}|} \left(\left| \frac{\partial g}{\partial u} \right|_{i,j}^{u_{i,j}^{\max}} + \left| \frac{\partial h}{\partial u} \right|_{i,j}^{u_{i,j}^{\max}} \right) \geq 0$$

where

$$\begin{aligned} \left| \frac{\partial g}{\partial u} \right|_{i,j}^{\max} &= \sup_{\substack{\tilde{u} \in [u_{i-1/2,j}^*, u_{i+1/2,j}^*] \\ \tilde{u} \in [u_{i-1/2,j}^*, u_{i+1/2,j}^*]}} \left(\frac{\partial g}{\partial u}(\tilde{u}, u_{i+1/2,j}^*) \right. \\ &\quad \left. - \frac{\partial g}{\partial u}(\tilde{u}, u_{i-1/2,j}^*) \right) \geq 0 \\ \left| \frac{\partial h}{\partial u} \right|_{i,j}^{\max} &= \sup_{\substack{\tilde{u} \in [u_{i,j-1/2}^*, u_{i,j+1/2}^*] \\ \tilde{u} \in [u_{i,j-1/2}^*, u_{i,j+1/2}^*]}} \left(\frac{\partial h}{\partial u}(\tilde{u}, u_{i,j+1/2}^*) \right. \\ &\quad \left. - \frac{\partial h}{\partial u}(\tilde{u}, u_{i,j-1/2}^*) \right) \geq 0 \end{aligned}$$

Many limiter functions satisfy the technical conditions (70) of Theorem 26. Some examples include

- the van Leer limiter

$$\Psi^{VL}(R) = \frac{R + |R|}{1 + |R|}$$

- the van Albada limiter

$$\Psi^{VA}(R) = \frac{R + R^2}{1 + R^2}$$

In addition, Koren (1988) has constructed the limiter

$$\Psi^K(R) = \frac{R + 2R^2}{2 - R + 2R^2}$$

which also satisfies the technical conditions (70) and corresponds for smooth solutions in 1-D to the most accurate $\kappa = 1/3$ MUSCL scheme of van Leer.

3.2.2 FV schemes on unstructured meshes utilizing linear reconstruction

Higher-order finite volume extensions of Godunov discretization to unstructured meshes are of the general form

$$\frac{du_j}{dt} = - \frac{1}{|T_j|} \sum_{v \in \partial T_j} g_{jk}(u_j, u_k), \quad \forall T_j \in \mathcal{T} \quad (71)$$

with the numerical flux $g_{jk}(u, v)$ given by the quadrature rule

$$g_{jk}(u_j^*, u_k^*) = \sum_{q=1}^Q \omega_q g(v_{jk}(x_q); u_j^*(x_q), u_k^*(x_q)) \quad (72)$$

where $\omega_q \in \mathbb{R}$ and $x_q \in e_{jk}$ represent quadrature weights and locations, $q = 1, \dots, Q$. Given the global space of piecewise constant cell averages, $u_h \in V_h^0$, the extrapolated states $u_{jk}^*(x)$ and $u_k^*(x)$ are evaluated using a p th order polynomial reconstruction operator, $R_p^0: V_h^0 \mapsto V_h^p$,

$$\begin{aligned} u_{jk}^-(x) &\equiv \lim_{\epsilon \downarrow 0} R_p^0(x - \epsilon v_{jk}(x); u_h) \\ u_{jk}^+(x) &\equiv \lim_{\epsilon \downarrow 0} R_p^0(x + \epsilon v_{jk}(x); u_h) \end{aligned}$$

for any $x \in e_{jk}$. In addition, it is assumed that the reconstruction satisfies the property $\frac{1}{|T_j|} \int_{T_j} R_p^0(x; u_h) dx = u_j$ stated previously in (64b). In the general finite volume formulation, the control volume shapes need not be convex; see, for example, Figure 4. Even so, the solution accuracy and maximum stable time step for explicit schemes may depend strongly on the shape of individual control volumes. In the special case of linear reconstruction, $R_p^0(x; u_h)$, the impact of control volume shape on stability of the scheme can be quantified more precisely. Specifically, the maximum principle analysis presented later for the scheme (71) reveals an explicit dependence on the geometrical shape

parameter

$$\Gamma^{\text{geom}} = \sup_{0 \leq \theta \leq 2\pi} \alpha^{-1}(\theta) \quad (73)$$

where $0 < \alpha(\theta) < 1$ represents the smallest fractional perpendicular distance from the gravity center to one of two minimally separated parallel hyperplanes with orientation θ and hyperplane location such that all quadrature points in the control volume lie between or on the hyperplanes as shown in Figure 5. Table 2 lists Γ^{geom} values for various control volume shapes in \mathbb{R}^1 , \mathbb{R}^2 , \mathbb{R}^3 , and \mathbb{R}^d . As might be expected, those geometries that have exact quadrature point symmetry with respect to the control volume gravity center have geometric shape parameters Γ^{geom} equal to 2 regardless of the number of space dimensions involved.

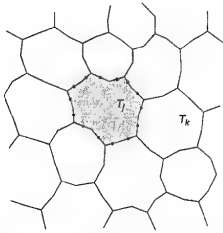


Figure 4. Polygonal control volume cell T_j and perimeter quadrature points (solid circles).

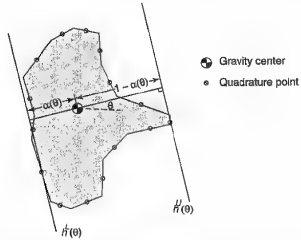


Figure 5. Minimally separated hyperplanes $h^l(\theta)$ and $h^u(\theta)$ and the fractional distance ratio $\alpha(\theta)$ for use in the calculation of Γ^{geom} .

Table 2. Reconstruction geometry factors for various control volume shapes utilizing midpoint quadrature rule.

| Control volume shape | Space dimension | Γ^{geom} |
|----------------------|-----------------|--|
| Segment | 1 | 2 |
| Triangle | 2 | 3 |
| Parallelogram | 2 | 2 |
| Regular n -gon | 2 | $n / \left\lceil \frac{n-1}{2} \right\rceil$ |
| Tetrahedron | 3 | 4 |
| parallelepiped | 3 | 2 |
| Simplex | d | $d+1$ |
| Hyper-parallelepiped | d | 2 |
| Polytope | d | Equation (73) |

Lemma 2 (Finite volume interval bounds on unstructured meshes, $R_1^0(x; u_h)$) The fully discrete finite volume scheme

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{|T_j|} \sum_{\nu_{jk} \in \partial T_j} g_{jk}(u_{jk}^{n-}, u_{jk}^{n+}), \quad \forall T_j \in \mathcal{T} \quad (74)$$

with monotone Lipschitz continuous numerical flux function, nonnegative quadrature weights, and linear reconstructions

$$u_{jk}^-(x) \equiv \lim_{\epsilon \downarrow 0} R_1^0(x - \epsilon \nu_{jk}(x); u_h), \quad x \in e_{jk}, \quad u_h \in V_h^0$$

$$u_{jk}^+(x) \equiv \lim_{\epsilon \downarrow 0} R_1^0(x + \epsilon \nu_{jk}(x); u_h), \quad x \in e_{jk}, \quad u_h \in V_h^0$$

with extremal trace values at control volume quadrature points

$$U_j^{\min} = \min_{\nu_{jk} \in \partial T_j} u_{jk}^-(x_q), \quad U_j^{\max} = \max_{\nu_{jk} \in \partial T_j} u_{jk}^+(x_q), \quad x_q \in e_{jk}, \quad 1 \leq q \leq Q$$

exhibits the local interpolated interval bound

$$\sigma_j U_j^{\min, n} + (1 - \sigma_j) u_j^n \leq u_j^{n+1} \leq (1 - \sigma_j) u_j^n + \sigma_j U_j^{\max, n} \quad (75)$$

with the time step proportional interpolation parameter σ_j defined by

$$\sigma_j \equiv \frac{\Delta t}{|T_j|} \Gamma^{\text{geom}} \sum_{\nu_{jk} \in \partial T_j} \sup_{\substack{\tilde{u} \in [U_j^{\min, n}, U_j^{\max, n}] \\ 1 \leq q \leq Q}} \left| \frac{\partial g_{jk}}{\partial u^+}(\nu_{jk}(x_q); \tilde{u}, \tilde{u}) \right| \quad (76)$$

that depends on the shape parameter Γ^{geom} defined in (73).

Given the two-sided bound of Lemma 2, a discrete maximum principle is obtained under a CFL-like time step restriction if the limits U_j^{\max} and U_j^{\min} can be bounded from

above and below respectively by the neighboring cell averages. This idea is given more precisely in the following theorem.

Theorem 27 (Finite volume maximum principle on unstructured meshes, R_1^0) Let u_j^{\min} and u_j^{\max} denote the minimum and maximum value of solution cell averages for a given cell T_j and corresponding adjacent cell neighbors, that is,

$$u_j^{\min} \equiv \min_{\nu_{jk} \in \partial T_j} (u_j, u_k) \quad \text{and} \quad u_j^{\max} \equiv \max_{\nu_{jk} \in \partial T_j} (u_j, u_k) \quad (77)$$

The fully discrete finite volume scheme

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{|T_j|} \sum_{\nu_{jk} \in \partial T_j} g_{jk}(u_{jk}^{n-}, u_{jk}^{n+}), \quad \forall T_j \in \mathcal{T} \quad (78)$$

with monotone Lipschitz continuous numerical flux function, nonnegative quadrature weights, and linear reconstructions

$$u_{jk}^-(x) \equiv \lim_{\epsilon \downarrow 0} R_1^0(x - \epsilon \nu_{jk}(x); u_h), \quad x \in e_{jk}, \quad u_h \in V_h^0$$

$$u_{jk}^+(x) \equiv \lim_{\epsilon \downarrow 0} R_1^0(x + \epsilon \nu_{jk}(x); u_h), \quad x \in e_{jk}, \quad u_h \in V_h^0$$

exhibits the local space-time maximum principle for each $T_j \in \mathcal{T}$

$$\min_{\nu_{jk} \in \partial T_j} (u_j^n, u_k^n) \leq u_j^{n+1} \leq \max_{\nu_{jk} \in \partial T_j} (u_j^n, u_k^n)$$

as well as the local spatial maximum principle at steady state ($u^{n+1} = u^n = u^*$)

$$\min_{\nu_{jk} \in \partial T_j} u_k^* \leq u_j^* \leq \max_{\nu_{jk} \in \partial T_j} u_k^*$$

if the linear reconstruction satisfies $\forall e_{jk} \in \partial T_j$ and $x_q \in e_{jk}$, $q = 1, \dots, Q$

$$\max(u_j^{\min, n}, u_k^{\min, n}) \leq u_{jk}^-(x_q) \leq \min(u_j^{\max, n}, u_k^{\max, n}) \quad (80)$$

under the time step restriction

$$1 - \frac{\Delta t}{|T_j|} \Gamma^{\text{geom}} \sum_{\nu_{jk} \in \partial T_j} \sup_{\substack{\tilde{u} \in [u_j^{\min, n}, u_j^{\max, n}] \\ 1 \leq q \leq Q}} \left| \frac{\partial g_{jk}}{\partial u^+}(\nu_{jk}(x_q); \tilde{u}, \tilde{u}) \right| \geq 0$$

with Γ^{geom} defined in (73).

Note that a variant of this theorem also holds if the definition of u_j^{\max} and u_j^{\min} are expanded to include more control volume neighbors. Two alternative definitions frequently

used when the control volume shape is a simplex are given by

$$u_j^{\min} \equiv \min_{T_k \in \mathcal{T}, T_j \cap T_k \neq \emptyset} u_k \quad \text{and} \quad u_j^{\max} \equiv \max_{T_k \in \mathcal{T}, T_j \cap T_k \neq \emptyset} u_k \quad (81)$$

These expanded definitions include adjacent cells whose intersection with T_j in \mathbb{R}^d need only be a set of measure zero or greater.

Slope limiters for linear reconstruction.

Given a linear reconstruction $R_1^0(x; u_h)$ that does not necessarily satisfy the requirements of Theorem 27, it is straightforward to modify the reconstruction so that the new modified reconstruction does satisfy the requirements of Theorem 27. For each control volume $T_j \in \mathcal{T}$, a modified reconstruction operator $\tilde{R}_1^0(x; u_h)$ of the form

$$\tilde{R}_1^0(x; u_h)|_{T_j} = u_j + \alpha_{T_j} (R_1^0(x; u_h)|_{T_j} - u_j)$$

is assumed for $\alpha_{T_j} \in [0, 1]$. By construction, this modified reconstruction correctly reproduces the control volume cell average for all values of α_{T_j} , that is,

$$\frac{1}{|T_j|} \int_{T_j} \tilde{R}_1^0(x; u_h) dx = u_j \quad (82)$$

The most restrictive value of α_{T_j} for each control volume T_j is then computed on the basis of the Theorem 27 constraint (80), that is,

$$\alpha_{T_j}^{\text{MM}} = \min_{\substack{\nu_{jk} \in \partial T_j \\ 1 \leq q \leq Q}} \begin{cases} \frac{\min(u_j^{\min, n}, u_k^{\min, n}) - u_j}{R_1^0(x_q; u_h)|_{T_j} - u_j} & \text{if } R_1^0(x_q; u_h)|_{T_j} > \min(u_j^{\min, n}, u_k^{\min, n}) \\ \frac{\max(u_j^{\max, n}, u_k^{\max, n}) - u_j}{R_1^0(x_q; u_h)|_{T_j} - u_j} & \text{if } R_1^0(x_q; u_h)|_{T_j} < \max(u_j^{\max, n}, u_k^{\max, n}) \\ 1 & \text{otherwise} \end{cases} \quad (83)$$

where u_j^{\max} and u_j^{\min} are defined in (77). When the resulting modified reconstruction operator is used in the extrapolation formulas (79), the discrete maximum principle of Theorem 27 is attained under a CFL-like time step restriction. By utilizing the inequalities

$$\max(u_j, u_k) \leq \min(u_j^{\max}, u_k^{\max})$$

and

$$\min(u_j, u_k) \geq \max(u_j^{\min}, u_k^{\min})$$

it is straightforward to construct a simpler but more restrictive limiter function

$$\alpha_{T_j}^{LM} = \min_{\substack{v_{qk} \in T_j \\ 1 \leq q \leq Q}} \begin{cases} \frac{\max(u_j, u_k) - u_j}{R_1^0(x_q; u_k)_{T_j} - u_j} & \text{if } R_1^0(x_q; u_k)_{T_j} > \max(u_j, u_k) \\ \frac{\min(u_j, u_k) - u_j}{R_1^0(x_q; u_k)_{T_j} - u_j} & \text{if } R_1^0(x_q; u_k)_{T_j} < \min(u_j, u_k) \\ 1 & \text{otherwise} \end{cases} \quad (84)$$

that yields modified reconstructions satisfying the technical conditions of Theorem 27. This simplified limiter (84) introduces additional slope reduction when compared to (83). This can be detrimental to the overall accuracy of the discretization. The limiter strategy (84) and other variants for simplicial control volumes are discussed further in Liu (1993), Wierse (1994), and Batten, Lambert and Causon (1996).

In Barth and Jespersen (1989), a variant of (83) was proposed

$$\alpha_{T_j}^{BJ} = \min_{\substack{v_{qk} \in T_j \\ 1 \leq q \leq Q}} \begin{cases} \frac{u_j^{\max} - u_j}{R_1^0(x_q; u_k)_{T_j} - u_j} & \text{if } R_1^0(x_q; u_k)_{T_j} > u_j^{\max} \\ \frac{u_j^{\min} - u_j}{R_1^0(x_q; u_k)_{T_j} - u_j} & \text{if } R_1^0(x_q; u_k)_{T_j} < u_j^{\min} \\ 1 & \text{otherwise} \end{cases} \quad (85)$$

Although this limiter function does not produce modified reconstructions satisfying the requirements of Theorem 27, using Lemma 2, it can be shown that the Barth and Jespersen limiter yields finite volume schemes (74) possessing a global extremum diminishing property, that is, that the solution maximum is nonincreasing and the solution minimum is nondecreasing between successive time levels. This limiter function produces the least amount of slope reduction when compared to the limiter functions (83) and (84). Note that in practical implementation, all three limiters (83), (84), and (85) require some modification to prevent near zero division for nearly constant solution data.

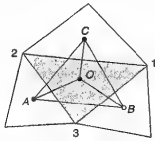


Figure 6. Triangle control volume Δ_{123} (shaded) with three adjacent cell neighbors.

3.2.3 Linear reconstruction operators on simplicial control volumes

Linear reconstruction operators on simplicial control volumes that satisfy the cell averaging requirement (64b) often exploit the fact that the cell average is also a pointwise value of any valid linear reconstruction evaluated at the gravity center of a simplex. This reduces the reconstruction problem to that of gradient estimation given pointwise samples at the gravity centers. In this case, it is convenient to express the reconstruction in the form

$$R_1^0(x; u_k)_{T_j} = u_j + (\nabla u_k)_{T_j} \cdot (x - x_j^g) \quad (86)$$

where x_j^g denotes the gravity center for the simplex T_j and $(\nabla u_k)_{T_j}$ is the gradient to be determined. Figure 6 depicts a 2-D simplex Δ_{123} and three adjacent neighboring simplices. Also shown are the corresponding four pointwise solution values $\{A, B, C, O\}$ located at gravity centers of each simplex. By selecting any three of the four pointwise solution values, a set of four possible gradients are uniquely determined, that is, $\{\nabla(ABC), \nabla(ABO), \nabla(BCO), \nabla(CAO)\}$. Using the example of Figure 6, a number of slope limited reconstruction techniques are possible for use in the finite volume scheme (78) that meet the technical conditions of Theorem 27.

1. Choose $(\nabla u_k)_{T_{123}} = \nabla(ABC)$ and limit the resulting reconstruction using (83) or (84). This technique is pursued in Barth and Jespersen (1989) but using the limiter (85) instead.
2. Limit the reconstructions corresponding to gradients $\nabla(ABC)$, $\nabla(ABO)$, $\nabla(BCO)$, and $\nabla(CAO)$ using (83) or (84) and choose the limited reconstruction with largest gradient magnitude. This technique is a generalization of that described in Batten, Lambert and Causon (1996) wherein limiter (84) is used.
3. Choose the unlimited reconstruction $\nabla(ABC)$, $\nabla(ABO)$, $\nabla(BCO)$, and $\nabla(CAO)$ with largest gradient magnitude that satisfies the maximum principle reconstruction bound inequality (80). If all reconstructions fail the bound inequality, the reconstruction gradient is set equal to zero; see Liu (1993).

3.2.4 Linear reconstruction operators on general control volumes shapes

In the case of linear reconstruction on general volume shapes, significant simplification is possible when compared with the general p -exact reconstruction formulation given in Section 3.2.5. It is again convenient to express the reconstruction in the form

$$R_1^0(x; u_k)_{T_j} = u_j + (\nabla u_k)_{T_j} \cdot (x - x_j^g) \quad (87)$$

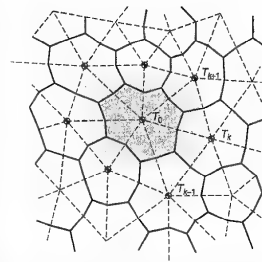


Figure 7. Triangulation of gravity center locations showing a typical control volume T_0 associated with the triangulation vertex v_0 with cyclically indexed graph neighbors $T_k, k = 1, \dots, N_0$.

where x_j^g denotes the gravity center for the control volume T_j and $(\nabla u_k)_{T_j}$ is the gradient to be determined. Two common techniques for simplified linear reconstruction include a simplified least squares technique and a Green–Gauss integration technique.

Simplified least squares linear reconstruction

As was exploited in the linear reconstruction techniques for simplicial control volumes, linear reconstructions satisfying (64b) on general control volume shapes are greatly simplified by exploiting the fact that the cell average value is also a pointwise value of all valid linear reconstructions evaluated at the gravity center of a general control volume shape. This again reduces the linear reconstruction problem to that of gradient estimation given pointwise values. In the simplified least squares reconstruction technique, a triangulation (2-D) or tetrahedralization (3-D) of gravity centers is first constructed as shown in Figure 7. Referring to this figure, for each edge of the simplex mesh incident to the vertex v_0 , an edge projected gradient constraint equation is constructed subject to a prespecified nonzero scaling w_k

$$w_k(\nabla u_k)_{T_0} \cdot (x_k^g - x_0^g) = w_k(u_k - u_0)$$

The number of edges incident to a simplex mesh vertex in \mathbb{R}^d is greater than or equal to d thereby producing the following generally nonsquare matrix of constraint equations

$$\begin{bmatrix} w_1 \Delta x_1^g & w_1 \Delta y_1^g \\ \vdots & \vdots \\ w_{N_0} \Delta x_{N_0}^g & w_{N_0} \Delta y_{N_0}^g \end{bmatrix} (\nabla u_k)_{T_0} = \begin{pmatrix} w_1(u_1 - u_0) \\ \vdots \\ w_{N_0}(u_{N_0} - u_0) \end{pmatrix}$$

or in abstract form

$$[\tilde{L}_1 \quad \tilde{L}_2] \nabla u = \tilde{f}$$

This abstract form can be symbolically solved in a least squares sense using an orthogonalization technique yielding the closed form solution

$$\nabla u = \frac{1}{l_{11}l_{22} - l_{12}^2} \begin{pmatrix} l_{22}(\tilde{L}_1 \cdot \tilde{f}) - l_{12}(\tilde{L}_2 \cdot \tilde{f}) \\ l_{11}(\tilde{L}_2 \cdot \tilde{f}) - l_{12}(\tilde{L}_1 \cdot \tilde{f}) \end{pmatrix} \quad (88)$$

with $l_{ij} = \tilde{L}_i \cdot \tilde{L}_j$. The form of this solution in terms of scalar dot products over incident edges suggests that the least squares linear reconstruction can be efficiently computed via an edge data structure without the need for storing a nonsquare matrix.

Green–Gauss linear reconstruction

This reconstruction technique specific to simplicial meshes assumes nodal solution values at vertices of the mesh which uniquely describes a C^0 linear interpolant, u_k . Gradients are then computed from the mean value approximation

$$(\nabla u_k)_{T_0} \approx \int_{\partial T_0} \nabla u_k \, dx = \int_{\partial T_0} u_k \, dv \quad (89)$$

For linear interpolants, the right-hand side term can be written in the following equivalent form using the configuration depicted in Figure 8

$$\int_{\partial T_0} \nabla u_k \, dx = \sum_{k=1}^{N_0} \frac{3}{2} (u_0 + u_k) v_{0k}$$

where v_{0k} represents any path integrated normal connecting pairwise adjacent simplex gravity centers, that is,

$$v_{0k} = \int_{x_0^{g+1/2}}^{x_k^{g+1/2}} dv \quad (90)$$

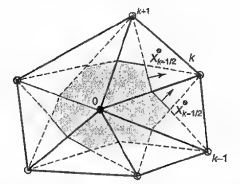


Figure 8. Median dual control volume T_0 demarcated by median segments of triangles incident to the vertex v_0 with cyclically indexed adjacent vertices $v_k, k = 1, \dots, N_0$.

A particularly convenient path is one that traces out portions of median segments as shown in Figure 8. These segments demarcate the so-called *median dual* control volume. By construction, the median dual volume $|T_0|$ is equal to $|\Omega_0|/(d+1)$ in \mathbb{R}^d . Consequently, a linear reconstruction operator on nonoverlapping median dual control volumes is given by

$$|T_0|(\nabla u_h)_{T_0} \approx \sum_{k=1}^{N_0} \frac{1}{2} (u_0 + u_k) v_{0k} \quad (91)$$

The gradient calculation is exact whenever the numerical solution varies linearly over the support of the reconstruction. Since mesh vertices are not located at the gravity centers of median dual control volumes, the cell averaging property (64b) and the bounds of Theorem 27 are only approximately satisfied using the Green–Gauss technique.

A number of slope limited linear reconstruction strategies for general control volume shapes are possible for use in the finite volume scheme (78) that satisfy the technical conditions of Theorem 27. Using the example depicted in Figure 7, let $\nabla_{k+1/2} u_h$ denote the unique linear gradient calculated from the cell average set $\{u_0, u_k, u_{k+1}\}$. Three slope limiting strategies that are direct counterparts of the simplex control volume case are

1. Compute $(\nabla u_h)_{T_0}$ using the least squares linear reconstruction or any other valid linear reconstruction technique and limit the resulting reconstruction using (83) or (84).
2. Limit the reconstructions corresponding to the gradients $\nabla_{k+1/2} u_h$, $k = 1, \dots, N_0$ and $(\nabla u_h)_{T_0}$ using (83) or (84) and choose the limited reconstruction with largest gradient magnitude.
3. Choose the unlimited reconstruction from $\nabla_{k+1/2} u_h$, $k = 1, \dots, N_0$ and $(\nabla u_h)_{T_0}$ with largest gradient magnitude that satisfies the maximum principle reconstruction bound inequality (80). If all reconstructions fail the bound inequality, the reconstruction gradient is set equal to zero.

3.2.5 General p -exact reconstruction operators on unstructured meshes

Abstractly, the reconstruction operator serves as a finite-dimensional pseudo inverse of the cell averaging operator A whose j th component A_j computes the cell average of the solution in T_j

$$A_j u = \frac{1}{|T_j|} \int_{T_j} u \, dx$$

The development of a general polynomial reconstruction operator, R_p^0 , that reconstructs p -degree polynomials from cell averages on unstructured meshes follows from the application of a small number of simple properties.

1. (Conservation of the mean) Given solution cell averages u_h , the reconstruction $R_p^0 u_h$ is required to have the correct cell average, that is,

$$\text{if } v = R_p^0 u_h \text{ then } u_h = Av$$

More concisely,

$$AR_p^0 = I$$

so that R_p^0 is a right inverse of the averaging operator A .

2. (p -exactness) A reconstruction operator R_p^0 is p -exact if $R_p^0 A$ reconstructs polynomials of degree p or less exactly. Denoting by \mathcal{P}_p the space of all polynomials of degree p ,

$$\text{if } u \in \mathcal{P}_p \text{ and } v = Au \text{ then } R_p^0 v = u$$

This can be written succinctly as

$$R_p^0 A|_{\mathcal{P}_p} = I$$

so that R_p^0 is a left inverse of the averaging operator A restricted to the space of polynomials of degree at most p .

3. (Compact support) The reconstruction in a control volume T_j should only depend on cell averages in a relatively small neighborhood surrounding T_j . Recall that a polynomial of degree p in \mathbb{R}^d contains $\binom{p+d}{d}$ degrees of freedom. The support set for T_j is required to contain at least this number of neighbors. As the support set becomes even larger for fixed p , not only does the computational cost increase, but eventually, the accuracy decreases as less valid data from further away is brought into the calculation.

Practical implementations of polynomial reconstruction operators fall into two classes:

- **Fixed support stencil reconstructions.** These methods choose a fixed support set as a preprocessing step. Various limiting strategies are then employed to obtain nonoscillatory approximation; see, for example, Barth and Frederickson (1990) and Delany (1996) for further details.
- **Adaptive support stencil reconstructions.** These ENO-like methods dynamically choose reconstruction stencils based on solution smoothness criteria; see, for

example, Harten and Chakravarthy (1991), Vankeersblick (1993), Abgrall (1994), Sonar (1997), and Sonar (1998) for further details.

3.2.6 Positive coefficient schemes on unstructured meshes

Several related positive coefficient schemes have been proposed on multidimensional simplicial meshes based on one-dimensional interpolation. The simplest example is the *upwind triangle scheme* as introduced by Billey *et al.* (1987), Desideri and Dervieux (1988) and Rostand and Stoufflet (1988) with later improved variants given by Jameson (1993) and Courède, Debiez and Dervieux (1998). These schemes are not Godunov methods in the sense that a single multidimensional gradient is not obtained in each control volume. The basis for these methods originates from the gradient estimation formula (91) generalized to the calculation of flux divergence on a median dual tessellation. In deriving this flux divergence formula, the assumption has been made that flux components vary linearly within a simplex yielding the discretization formula

$$\begin{aligned} \int_{T_j} \operatorname{div}(f) \, dx &= \int_{\partial T_j} f \cdot \nu \, dv \\ &= \sum_{\nu_{jk} \in \partial T_j} \frac{1}{2} (f(u_j) + f(u_k)) \cdot \nu_{jk} \end{aligned}$$

where ν_{jk} is computed from a median dual tessellation using (90). This discretization is the unstructured mesh counterpart of central differencing on a structured mesh. Schemes using this discretization of flux divergence lack sufficient stability properties for computing solutions of general nonlinear conservation laws. This lack of stability can be overcome by adding suitable diffusion terms. One of the simplest modifications is motivated by upwind domain of dependence arguments yielding the numerical flux

$$g_{jk}(u_j, u_k) = \frac{1}{2} (f(u_j) + f(u_k)) \cdot \nu_{jk} - \frac{1}{2} |a_{jk}| \Delta_{jk} u \quad (92)$$

with a_{jk} a mean value (a.k.a. Murman–Cole) linearization satisfying

$$\nu_{jk} \cdot \Delta_{jk} f = a_{jk} \Delta_{jk} u$$

Away from sonic points where $f'(u^*) = 0$ for $u^* \in [u_j, u_{j+1}]$, this numerical flux is formally an E-flux satisfying (28). With suitable modifications of a_{jk} near sonic points, it is then possible to produce a modified numerical flux that is an E-flux for all data; see Osher (1984). Theorems 5, 6, and 7 show that schemes such as (22) using E-fluxes exhibit local discrete maximum principles and L_∞ stability.

Unfortunately, schemes based on (92) are too dissipative for most practical calculations. The main idea in the upwind triangle scheme is to add antidiffusion terms to the numerical flux function (92) such that the sum total of added diffusion and antidiffusion terms in the numerical flux function vanish entirely whenever the numerical solution varies linearly over the support of the flux function. In all remaining situations, the precise amount of antidiffusion is determined from maximum principle analysis.

Theorem 28 (Maximum principles for the upwind triangle scheme) Let \mathcal{T} denote the median dual tessellation of an underlying simplicial mesh. Also, let u_j denote the nodal solution value at a simplex vertex in one-to-one dual correspondence with the control volume $T_j \in \mathcal{T}$ such that a C^0 linear solution interpolant is uniquely specified on the simplicial mesh. Let $g_{jk}(u_j, u_j, u_k, u_k)$ denote the numerical flux function with limiter function $\Psi(\cdot): \mathbb{R} \mapsto \mathbb{R}$

$$\begin{aligned} g_{jk}(u_j, u_j, u_k, u_k) &= \frac{1}{2} (f(u_j) + f(u_k)) \cdot \nu_{jk} \\ &\quad - \frac{1}{2} a_{jk}^+ \left(1 - \Psi \left(\frac{h_{jk} \Delta_{jk} f}{h_{jk} \Delta_{jk} u} \right) \right) \Delta_{jk} u \\ &\quad + \frac{1}{2} a_{jk}^- \left(1 - \Psi \left(\frac{h_{jk} \Delta_{jk} u}{h_{jk} \Delta_{jk} u} \right) \right) \Delta_{jk} u \end{aligned}$$

utilizing the mean value speed a_{jk} satisfying

$$\nu_{jk} \cdot \Delta_{jk} f = a_{jk} \Delta_{jk} u$$

and variable spacing parameter $h_{jk} = |\Delta_{jk} x|$. The fully discrete finite volume scheme

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{|T_j|} \sum_{\nu_{jk} \in \partial T_j} g_{jk}(u_j^n, u_j^n, u_k^n, u_k^n), \quad \forall T_j \in \mathcal{T}$$

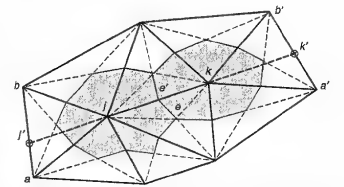


Figure 9. Triangle complex used in the upwind triangle schemes showing the linear extension of e_{jk} into neighboring triangle for the determination of points x_j' and x_k' .

with linearly interpolated values u_j and u_k as depicted in Figure 9 exhibits the local space–time maximum principle

$$\min_{v_{0k} \in \mathcal{T}_j} (u_j^n, u_k^n) \leq u_j^{n+1} \leq \max_{v_{0k} \in \mathcal{T}_j} (u_j^n, u_k^n)$$

and the local spatial maximum principle at steady state ($u^{n+1} = u^n = u^*$)

$$\min_{v_{0k} \in \mathcal{T}_j} u_k^* \leq u_j^* \leq \max_{v_{0k} \in \mathcal{T}_j} u_k^*$$

if the limiter $\Psi(R)$ satisfies $\forall R \in \mathbb{R}$

$$0 \leq \frac{[\Psi(R)]}{R}, \quad 0 \leq \Psi(R) \leq 2$$

Some standard limiter functions that satisfy the requirements of Theorem 28 include

- the MinMod limiter with maximum compression parameter equal to 2

$$\Psi^{\text{MM}}(R) = \max(0, \min(R, 2))$$

- the van Leer limiter

$$\Psi^{\text{VL}}(R) = \frac{R + |R|}{1 + |R|}$$

Other limiter formulations involving three successive one-dimensional slopes are given in Jameson (1993) and Courbade, Debiez and Dervieux (1998).

4 FURTHER ADVANCED TOPICS

The remainder of this chapter will consider several extensions of the finite volume method. Section 4.1 considers a class of higher-order accurate discretizations in time that still preserve the stability properties of the fully discrete schemes using Euler time integration. This is followed by a discussion of generalizations of the finite volume method for problems including second-order diffusion terms and the extension to systems of nonlinear conservation laws.

4.1 Higher-order time integration schemes

The derivation of finite volume schemes in Section 2 began with a semidiscrete formulation (21) that was later extended to a fully discrete formulation (22) by the introduction of first-order accurate forward Euler time integration. These latter schemes were then subsequently extended to higher-order accuracy in space using a variety of techniques. For

many computing problems of interest, first-order accuracy in time is then no longer enough. To overcome this low-order accuracy in time, a general class of higher-order accurate time integration methods was developed that preserve stability properties of the fully discrete scheme with forward Euler time integration. Following Gottlieb, Shu and Tadmor (2001) and Shu (2002), these methods will be referred to as *strong stability preserving* (SSP) time integration methods.

Explicit SSP Runge–Kutta methods were originally developed by Shu (1988), Shu and Osher (1988) and Gottlieb and Shu (1998) and called *TVD Runge–Kutta time discretizations*. In a slightly more general approach, total variation bounded (TVB) Runge–Kutta methods were considered by Cockburn and Shu (1989), Cockburn, Lin and Shu (1989), Cockburn, Hou and Shu (1990) and Cockburn and Shu (1998) in combination with the discontinuous Galerkin discretization in space. K  tner (2000) later gave error estimates for second-order TVD Runge–Kutta finite volume approximations of hyperbolic conservation laws.

To present the general framework of SSP Runge–Kutta methods, consider writing the semidiscrete finite volume method in the following form

$$\frac{d}{dt} U(t) = L(U(t)) \quad (93)$$

where $U = U(t)$ denotes the solution vector of the semidiscrete finite volume method. Using this notation together with forward Euler time integration yields the fully discrete form

$$U^{n+1} = U^n - \Delta t L(U^n) \quad (94)$$

where U^n is now an approximation of $U(t^n)$. As demonstrated in Section 2.2, the forward Euler time discretization is stable with respect to the L^∞ -norm, that is,

$$\|U^{n+1}\|_\infty \leq \|U^n\|_\infty \quad (95)$$

subject to a CFL-like time step restriction

$$\Delta t \leq \Delta t_0 \quad (96)$$

With this assumption, a time integration method is said to be SSP (see Gottlieb, Shu and Tadmor, 2001) if it preserves the stability property (95), albeit with perhaps a slightly different restriction on the time step

$$\Delta t \leq c \Delta t_0 \quad (97)$$

where c is called the CFL coefficient of the SSP method. In this framework, a general objective is to find SSP methods that are higher-order accurate, have low computational cost

and storage requirements, and have preferably a large CFL coefficient. Note that the TVB Runge–Kutta methods can be embedded into this class if the following relaxed notion of stability is assumed

$$\|U^{n+1}\|_\infty \leq (1 + C(\Delta t)) \|U^n\|_\infty \quad (98)$$

4.1.1 Explicit SSP Runge–Kutta methods

Following Shu and Osher (1988) and the review articles by Gottlieb, Shu and Tadmor (2001) and Shu (2002), a general m stage Runge–Kutta method for integrating (93) in time can be algorithmically represented as

$$\begin{aligned} \tilde{U}^0 &:= U^n \\ \tilde{U}^l &:= \sum_{k=0}^{l-1} (\alpha_{lk} \tilde{U}^k + \beta_{lk} \Delta t L(\tilde{U}^k)), \\ \alpha_{lk} &\geq 0, \quad l = 1, \dots, m \\ U^{n+1} &:= \tilde{U}^m \end{aligned} \quad (99)$$

To ensure consistency, the additional constraint $\sum_{k=0}^{l-1} \alpha_{lk} = 1$ is imposed. If, in addition, all β_{lk} are assumed to be non-negative, it is straightforward to see that the method can be written as a convex (positive weighted) combination of simple forward Euler steps with Δt replaced by $(\beta_{lk}/\alpha_{lk})\Delta t$. From this property, Shu and Osher (1988) concluded the following lemma:

Lemma 3. *If the forward Euler method (94) is L^∞ -stable subject to the CFL condition (96), then the Runge–Kutta method (99) with $\beta_{lk} \geq 0$ is SSP, that is, the method is L^∞ -stable under the time step restriction (97) with CFL coefficient*

$$c = \min_{l,k} \frac{\beta_{lk}}{\alpha_{lk}} \quad (100)$$

In the case of negative β_{lk} , a similar result can be proven; see (Shu and Osher, 1988).

4.1.2 Optimal second- and third-order nonlinear SSP Runge–Kutta methods

Gottlieb, Shu and Tadmor (2001) (Proposition 3.1) show that the maximal CFL coefficient for any m -stage, m th order accurate SSP Runge–Kutta methods is $c = 1$. Therefore, SSP Runge–Kutta methods that achieve $c = 1$ are termed ‘optimal’. Note that this restriction is not true if the number of stages is higher than the order of accuracy; see Shu (1988).

Optimal second- and third-order nonlinear SSP Runge–Kutta methods are given in Shu and Osher (1988). The

optimal second-order, two-stage nonlinear SSP Runge–Kutta method is given by

$$\begin{aligned} \tilde{U}^0 &:= U^n \\ \tilde{U}^1 &:= \tilde{U}^0 + \Delta t L(\tilde{U}^0) \\ U^{n+1} &:= \frac{1}{2} \tilde{U}^0 + \frac{1}{2} \tilde{U}^1 + \frac{1}{2} \Delta t L(\tilde{U}^1) \end{aligned}$$

This method corresponds to the well-known method of Heun. Similarly, the optimal third-order, three-stage nonlinear SSP Runge–Kutta method is given by

$$\begin{aligned} \tilde{U}^0 &:= U^n \\ \tilde{U}^1 &:= \tilde{U}^0 + \Delta t L(\tilde{U}^0) \\ \tilde{U}^2 &:= \frac{3}{4} \tilde{U}^0 + \frac{1}{4} \tilde{U}^1 + \frac{1}{4} \Delta t L(\tilde{U}^1) \\ U^{n+1} &:= \frac{1}{6} \tilde{U}^0 + \frac{4}{6} \tilde{U}^2 + \frac{1}{6} \Delta t L(\tilde{U}^2) \end{aligned}$$

Further methods addressing even higher-order accuracy or lower storage requirements are given in the review articles of Gottlieb, Shu and Tadmor (2001) and Shu (2002) where SSP multistep methods are also discussed.

4.2 Discretization of elliptic problems

Finite volume methods for elliptic boundary value problems have been proposed and analyzed under a variety names: box methods, covolume methods, diamond cell methods, integral finite difference methods, and finite volume element (FVE) methods; see Bank and Rose (1987), Cai (1991), S  li (1991), Lazarov, Michev and Vassilevsky (1996), Viozat *et al.* (1998), Chatzipantelidis (1999), Chou and Li (2000), Hermeline (2000), Eymard, Gallu  t and Herbin (2000), and Ewing, Lin and Lin (2002). These methods address the discretization of the following standard elliptic problem in a convex polygonal domain $\Omega \subset \mathbb{R}^2$

$$\begin{aligned} -\nabla \cdot A \nabla u &= f \quad \text{in } \Omega \\ u(x) &= 0 \quad \text{on } \partial\Omega \end{aligned} \quad (101)$$

for $A \in \mathbb{R}^{2 \times 2}$, a symmetric positive definite matrix (assumed constant). Provided $f \in H^0(\Omega)$, then a solution u exists such that $u \in H_0^{1/2}(\Omega)$, $-1 \leq \beta \leq 1$, $\beta \neq \pm 1/2$, where $H^1(\Omega)$ denotes the Sobolev space of order s in Ω .

Nearly all the above mentioned methods can be recast in Petrov–Galerkin form using a piecewise constant test space together with a conforming trial space. A notable exception is given in Chatzipantelidis (1999) wherein nonconforming Crouzeix–Raviart elements are utilized and analyzed. To formulate and analyze the Petrov–Galerkin representation, two tessellations of Ω are considered: a triangulation \mathcal{T}

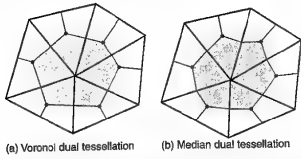


Figure 10. Two control volume variants used in the finite volume discretization of second-order derivative terms (a) Voronoi dual where edges of the Voronoi dual are perpendicular to edges of the triangulation and (b) median dual formed from median dual segments in each triangle.

with simplicial elements $K \in \mathcal{T}$ and a dual tessellation \mathcal{T}^* with control volumes $T \in \mathcal{T}^*$. In the class of conforming trial space methods such as the FVE method, a globally continuous, piecewise p th order polynomial trial space with zero trace value on the physical domain boundary is constructed

$$X_h = \{v \in C^0(\Omega) \mid v|_K \in \mathcal{P}_p(K), \forall K \in \mathcal{T} \text{ and } v|_{\partial\Omega} = 0\}$$

using nodal Lagrange elements on the simplicial mesh. A dual tessellation \mathcal{T}^* of the Lagrange element is then constructed; see, for example, Figure 10, which shows a linear Lagrange element with two dual tessellation possibilities. These dual tessellated regions form control volumes for the finite volume method. The tessellation technique extends to higher-order Lagrange elements in a straightforward way. A piecewise constant test space is then constructed using \mathcal{T}^*

$$Y_h = \{v \mid v|_T \in \chi(T), \forall T \in \mathcal{T}^*\}$$

where $\chi(T)$ is a characteristic function in the control volume T . The finite volume element discretization of (101) then yields the following Petrov–Galerkin formulation: Find $u_h \in X_h$ such that

$$\sum_{T \in \mathcal{T}^*} \left(\int_T w_h A \nabla u_h \cdot \nu + \int_T w_h f \, dx \right) = 0, \quad \forall w_h \in Y_h \quad (102)$$

The analysis of (102) by Ewing, Lin and Lin (2002) using linear elements gives an a priori estimate in an L^2 norm that requires the least amount of solution regularity when compared to previous methods of analysis.

Theorem 29 (FVE a priori error estimate, Ewing, Lin and Lin, 2002) Assume a 2-D quasi-uniform triangulation

\mathcal{T} with dual tessellation \mathcal{T}^* such that $\exists C > 0$ satisfying

$$C^{-1}h^2 \leq |T| \leq Ch^2, \quad \forall T \in \mathcal{T}^*$$

Assume that u and u_h are solutions of (101) and (102) respectively with $u \in H^2(\Omega)$, $f \in H^\beta$, $(0 \leq \beta \leq 1)$. Then $\exists C' > 0$ such that the a priori estimate holds

$$\|u - u_h\|_{L^2(\Omega)} \leq C' (h^2 \|u\|_{H^2(\Omega)} + h^{1+\beta} \|f\|_{H^\beta(\Omega)}) \quad (103)$$

Unlike the finite element method, the error estimate (103) reveals that optimal order convergence is obtained only if $f \in H^\beta$ with $\beta \geq 1$. Moreover, numerical results show that the source term regularity cannot be reduced without deteriorating the measured convergence rate. Optimal convergence rates are also shown for the nonconforming Crouzeix–Raviart element based finite volume method analyzed by Chatzipantelidis (1999) for $u \in H^2(\Omega)$ and $f \in H^1(\Omega)$.

An extensive presentation and analysis of finite volume methods for elliptic equations without utilizing a Petrov–Galerkin formulation is given in Eymard, Gallouët and Herbin (2000). In this work, general boundary conditions that include nonhomogeneous Dirichlet, Neumann, Robin conditions are discussed. In addition, the analysis is extended to general elliptic problems in divergence form, including convection, reaction, and singular source terms.

4.3 Conservation laws including diffusion terms

As demonstrated in Section 1, hyperbolic conservation laws are often approximations to physical problems with small or nearly vanishing viscosity. In other problems, the quantitative solution effects of these small viscosity terms are actually sought. Consequently, it is necessary in these problems, to include viscosity terms into the conservation law formulation. As a model for these latter problems, a second-order Laplacian term with small diffusion parameter is added to the first-order Cauchy problem, that is,

$$\partial_t u + \nabla \cdot f(u) - \varepsilon \Delta u = 0 \quad \text{in } \mathbb{R}^d \times \mathbb{R}^+ \quad (104a)$$

$$u(x, 0) = u_0 \quad \text{in } \mathbb{R}^d \quad (104b)$$

Here, $u(x, t): \mathbb{R}^d \times \mathbb{R}^+ \rightarrow \mathbb{R}$ denotes the dependent solution variable, $f \in C(\mathbb{R})$ the hyperbolic flux, and $\varepsilon \geq 0$ a small diffusion coefficient. Application of the divergence and Gauss theorems to (104a) integrated in a region T yields the following integral conservation law form

$$\frac{\partial}{\partial t} \int_T u \, dx + \int_{\partial T} f(u) \cdot \nu - \int_{\partial T} \varepsilon \nabla u \cdot \nu = 0 \quad (105)$$

A first goal is to extend the fully discrete form (22) of Section 2 to the integral conservation law (105) by the introduction of a numerical diffusion flux function $d_{jk}(u_h)$ for a control volume $T_j \in \mathcal{T}$ such that

$$\int_{\partial T_j} \varepsilon \nabla u \cdot \nu \approx \sum_{e_{jk} \in \partial T_j} d_{jk}(u_h)$$

When combined with the general finite volume formulation (22) for hyperbolic conservation laws, the following fully discrete scheme is produced

$$u_j^{n+1} = u_j^n - \frac{\Delta t^n}{|T_j|} \sum_{e_{jk} \in \partial T_j} (g_{jk}(u_j^n, u_k^n) - d_{jk}(u_h^n)), \quad \forall T_j \in \mathcal{T} \quad (106)$$

In this equation, the index m may be chosen either as n or $n+1$, corresponding to an explicit or implicit discretization.

4.3.1 Choices of the numerical diffusion flux d_{jk}

The particular choice of the numerical diffusion flux function d_{jk} depends on the type of control volume that is used. Since the approximate solution u_h is assumed to be a piecewise constant function, the definition of d_{jk} involves a gradient reconstruction of u_h in the normal direction to each cell interface e_{jk} . The reconstruction using piecewise constant gradients is relatively straightforward if the control volumes are vertex-centered, or if the cell interfaces are perpendicular to the straight lines connecting the storage locations (see Figure 10).

Vertex-centered finite volume schemes.

In the case of vertex-centered control volumes such as the median dual control volume, a globally continuous, piecewise linear approximate solution \tilde{u}_h is first reconstructed on the primal mesh. $\nabla \tilde{u}_h$ is then continuous on the control volume interfaces and the numerical diffusion flux straightforwardly computed as

$$d_{jk}(u_h^n) = \int_{e_{jk}} \nabla \tilde{u}_h^n \cdot \nu_{jk} \quad (107)$$

Cell-centered finite volume schemes.

In the case of cell-centered finite volume schemes, where an underlying primal–dual mesh relationship may not exist, a simple numerical diffusion flux can be constructed whenever cell interfaces are exactly or approximately perpendicular to the straight lines connecting the storage locations, for example, Voronoi meshes, quadrilateral meshes, and so on. In these cases, the reconstructed gradient of u_h projected normal to the cell interface e_{jk} can be represented

by

$$\nabla u_h^n \cdot \nu_{jk} = \frac{u_k^n - u_j^n}{|x_k - x_j|}$$

where x_i denotes the storage location of cell T_i . The numerical diffusion flux for this case is then given by

$$d_{jk}(u_h^n) = \frac{|e_{jk}|}{|x_k - x_j|} (u_k^n - u_j^n) \quad (108)$$

Further possible constructions and generalizations are given in Eymard, Gallouët and Herbin (2001), Gallouët, Herbin and Vignal (2000), and Herbin and Ohlberger (2002).

4.3.2 Note on stability, convergence, and error estimates

Stability analysis reveals a CFL-like stability condition for the explicit scheme choice ($m = n$) in (106)

$$\Delta t^n \leq \frac{\alpha^2 (h_{\min}^n)^2}{\alpha L_\varepsilon h_{\min}^n + \varepsilon}$$

where L_ε denotes the Lipschitz constant of the hyperbolic numerical flux, α is a positive mesh dependent parameter, and ε is the diffusion coefficient. In constructing this bound, a certain form of shape regularity is assumed such that there exists an $\alpha > 0$ such that for all j, k with $h_k = \text{diam}(T_k)$

$$\alpha h_k^2 \leq |T_k|, \quad \alpha |\partial T_k| \leq h_k, \quad \alpha h_k \leq |x_k - x_j| \quad (109)$$

Thus, Δt^n is of the order h^2 for large ε and of the order h for $\varepsilon \leq h$. In cases where the diffusion coefficient is larger than the mesh size, it is advisable to use an implicit scheme ($m = n+1$). In this latter situation, no time step restriction has to be imposed; see Eymard *et al.* (2002) and Ohlberger (2001b).

In order to demonstrate the main difficulties when analyzing convection-dominated problems, consider the following result from Feistauer *et al.* (1999) for a homogeneous diffusive boundary value problem. In this work, a mixed finite volume, finite element method sharing similarities with the methods described above is used to obtain a numerical approximation u_h of the exact solution u . Using typical energy-based techniques, they prove the following a priori error bound.

Theorem 30. For initial data $u_0 \in L^\infty(\mathbb{R}^2) \cap W^{1,2}(\mathbb{R}^2)$ and $\tau > 0$ there exist constants $c_1, c_2 > 0$ independent of ε such that

$$\|u(\cdot, t^n) - u_h(\cdot, t^n)\|_{L^2(\Omega)} \leq c_1 h \varepsilon^{c_2/\sqrt{\varepsilon}} \quad (110)$$

This estimate is fundamentally different from estimates for the purely hyperbolic problems of Sections 2 and 3. Specifically, this result shows how the estimate strongly depends on the small parameter ε ; ultimately becoming unbounded as ε tends to zero.

In the context of convection dominated or degenerate parabolic equations, Kruzkov-techniques have been recently used by Carrillo (1999) and Karlsen and Risebro (2000) in proving uniqueness and stability of solutions. Utilizing these techniques, convergence of finite volume schemes (uniform with respect to $\varepsilon \rightarrow 0$) was proven in Eymard *et al.* (2002) and a priori error estimates were obtained for viscous approximations in Evje and Karlsen (2002) and Eymard, Gallouët and Herbin (2002). Finally, in Ohlberger (2001a, 2001b) uniform a posteriori error estimates suitable for adaptive meshing are given. Using the theory of nonlinear semigroups, continuous dependence results were also obtained in Cockburn and Gripenberg (1999) (see also Cockburn (2003) for a review).

4.4 Extension to systems of nonlinear conservation laws

A positive attribute of finite volume methods is the relative ease with which the numerical discretization schemes of Sections 2 and 3 can be algorithmically extended to systems of nonlinear conservation laws of the form

$$\partial_t u + \nabla \cdot f(u) = 0 \quad \text{in } \mathbb{R}^d \times \mathbb{R}^+ \quad (111a)$$

$$u(x, 0) = u_0(x) \quad \text{in } \mathbb{R}^d \quad (111b)$$

where $u(x, t): \mathbb{R}^d \times \mathbb{R}^+ \rightarrow \mathbb{R}^m$ denotes the vector of dependent solution variables, $f(u): \mathbb{R}^m \rightarrow \mathbb{R}^{m \times d}$ denotes the flux vector, and $u_0(x): \mathbb{R}^d \rightarrow \mathbb{R}^m$ denotes the initial data vector at time $t = 0$. It is assumed that this system is strictly hyperbolic, that is, the eigenvalues of the flux Jacobian $A(v; u) = \partial f / \partial u \cdot v$ are real and distinct for all bounded $v \in \mathbb{R}^d$.

The main task in extending finite volume methods to systems of nonlinear conservation laws is the construction of a suitable numerical flux function. To gain insight into this task, consider the one-dimensional linear Cauchy problem for $u(x, t): \mathbb{R} \times \mathbb{R}^+ \rightarrow \mathbb{R}^m$ and $u_0(x): \mathbb{R} \rightarrow \mathbb{R}^m$

$$\begin{aligned} \partial_t u + \partial_x (Au) &= 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^+ \\ u(x, 0) &= u_0(x) \quad \text{in } \mathbb{R} \end{aligned} \quad (112)$$

where $A \in \mathbb{R}^{m \times m}$ is a constant matrix. Assume the matrix A has m real and distinct eigenvalues, $\lambda_1 < \lambda_2 < \dots < \lambda_m$, with corresponding right and left eigenvectors denoted by $r_k \in \mathbb{R}^m$ and $l_k \in \mathbb{R}^m$ respectively for $k = 1, \dots, m$.

Furthermore, let $X \in \mathbb{R}^{m \times m}$ denote the matrix of right eigenvectors, $X = [r_1, \dots, r_m]$, and $\Lambda \in \mathbb{R}^{m \times m}$ the diagonal matrix of eigenvalues, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$ so that $A = X\Lambda X^{-1}$. The one-dimensional system (112) is readily decoupled into scalar equations via the transformation into characteristic variables $\alpha = X^{-1}u$ for $\alpha \in \mathbb{R}^m$

$$\begin{aligned} \partial_t \alpha + \partial_x (\Lambda \alpha) &= 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^+ \\ \alpha(x, 0) &= \alpha_0(x) \quad \text{in } \mathbb{R} \end{aligned} \quad (113)$$

and component-wise solved exactly

$$\alpha^{(k)}(x, t) = \alpha_0^{(k)}(x - \lambda_k t), \quad k = 1, \dots, m$$

or recombined in terms of the original variables

$$u(x, t) = \sum_{k=1}^m l_k \cdot u_0(x - \lambda_k t) r_k$$

Using this solution, it is straightforward to solve exactly the associated Riemann problem for $w(\xi, \tau) \in \mathbb{R}^m$

$$\partial_t w + \partial_\xi (Aw) = 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^+$$

with initial data

$$w(\xi, 0) = \begin{cases} u & \text{if } \xi < 0 \\ v & \text{if } \xi > 0 \end{cases}$$

thereby producing the following Godunov-like numerical flux function

$$\begin{aligned} g(u, v) &= Aw(\tau, 0) \\ &= \frac{1}{2}(Au + Av) - \frac{1}{2}|A|(v - u) \end{aligned} \quad (114)$$

with $|A| \equiv X|\Lambda|X^{-1}$. When used in one-dimensional discretization together with piecewise constant solution representation, the linear numerical flux (114) produces the well-known Courant–Isaacson–Rees (CIR) upwind scheme for linear systems of hyperbolic equations

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left(A^+ (u_j^n - u_{j-1}^n) + A^- (u_{j+1}^n - u_j^n) \right)$$

where $A^\pm \equiv X\Lambda^\pm X^{-1}$. Note that higher-order accurate finite volume methods with slope limiting procedures formally extend to this linear system via component-wise slope limiting of the characteristic components $\alpha^{(k)}$, $k = 1, \dots, m$ for use in the numerical flux (114).

4.4.1 Numerical flux functions for systems of conservation laws

In Godunov's original work (see Godunov, 1959), exact solutions of the one-dimensional nonlinear Riemann problem of gas dynamics were used in the construction of a similar numerical flux function

$$g^G(u, v) = f(w(0, t_+)) \cdot v \quad (115)$$

where $w(\xi, \tau) \in \mathbb{R}^m$ is now a solution of a nonlinear Riemann problem

$$\partial_t w + \partial_\xi f(w) = 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^+$$

with initial data

$$w(\xi, 0) = \begin{cases} u & \text{if } \xi < 0 \\ v & \text{if } \xi > 0 \end{cases}$$

Recall that solutions of the Riemann problem for gas dynamic systems are a composition of shock, contact, and rarefaction wave family solutions. For the gas dynamic equations considered by Godunov, a unique solution of the Riemann problem exists for general states u and v except those states producing a vacuum. Even so, the solution of the Riemann problem is both mathematically and computationally nontrivial. Consequently, a number of alternative numerical fluxes have been proposed that are more computationally efficient. These alternative numerical fluxes can be sometimes interpreted as approximate Riemann solvers. A partial list of alternative numerical fluxes is given here. A more detailed treatment of this subject is given in Godlewski and Raviart (1991), Kröner (1997), and LeVeque (2002).

• **Osher–Solomon flux** (Osher and Solomon, 1982). This numerical flux is a system generalization of the Enquist–Osher flux of Section 2. All wave families are approximated in state space as rarefaction or inverted rarefaction waves with Lipschitz continuous partial derivatives. The Osher–Solomon numerical flux is of the form

$$g^{OS}(u, v) = \frac{1}{2}(f(u) + f(v)) \cdot v - \frac{1}{2} \int_u^v |A(v; w)| dw$$

where $|A|$ denotes the usual matrix absolute value. By integrating on m rarefaction wave integral subpaths that are each parallel to a right eigenvector, a system decoupling occurs on each subpath integration. Furthermore, for the gas dynamic equations with ideal gas law, it is straightforward to construct $m - 1$ Riemann invariants on each subpath thereby eliminating

the need for path integration altogether. This reduces the numerical flux calculation to purely algebraic computations with special care taken at sonic points; see Osher and Solomon (1982).

• **Roe flux** (Roe, 1981). Roe's numerical flux can be interpreted as approximating all wave families as discontinuities. The numerical flux is of the form

$$\begin{aligned} g^{Roe}(u, v) &= \frac{1}{2}(f(u) + f(v)) \cdot v \\ &\quad - \frac{1}{2}|A(v; u, v)|(v - u) \end{aligned}$$

where $A(v; u, v)$ is the 'Roe matrix' satisfying the matrix mean value identity

$$(f(v) - f(u)) \cdot v = A(v; u, v)(v - u)$$

with $A(v; u, u) = A(v; u)$. For the equations of gas dynamics with ideal gas law, the Roe matrix takes a particularly simple form. Steady discrete mesh-aligned shock profiles are resolved with one intermediate point. The Roe flux does not preclude the formation of entropy violating expansion shocks unless additional steps are taken near sonic points.

• **Steger–Warming flux vector splitting** (Steger and Warming, 1981). Steger and Warming considered a splitting of the flux vector for the gas dynamic equations with ideal gas law that exploited the fact that the flux vector is homogeneous of degree one in the conserved variables. From this homogeneity property, Euler's identity then yields that $f(u) \cdot v = A(v; u)u$. Steger and Warming then considered the matrix splitting

$$A = A^+ + A^-, \quad A^\pm \equiv X\Lambda^\pm X^{-1}$$

where Λ^\pm is computed component-wise. From this matrix splitting, the final upwind numerical flux function was constructed as

$$g^{SW}(u, v) = A^+(v; u)u + A^-(v; v)v$$

Although not part of their explicit construction, for the gas dynamic equations with ideal gas law, the Jacobian matrix $\partial g^{SW} / \partial u$ has eigenvalues that are all nonnegative and the Jacobian matrix $\partial g^{SW} / \partial v$ has eigenvalues that are all nonpositive whenever the ratio of specific heats γ lies in the interval $[1, 5/3]$. The matrix splitting leads to numerical fluxes that do not vary smoothly near sonic and stagnation points. Use of the Steger–Warming flux splitting in the schemes of Sections 2 and 3 results in rather poor resolution

of linearly degenerate contact waves and velocity slip surfaces due to the introduction of excessive artificial diffusion for these wave families.

- **Van Leer flux vector splitting.** Van Leer (1982) provided an alternative flux splitting for the gas dynamic equations that produces a numerical flux of the form

$$g^{VL}(u, v) = f^-(u) + f^+(v)$$

using special Mach number polynomials to construct fluxes that remain smooth near sonic and stagnation points. As part of the splitting construction, the jacobian matrix $\partial g^{SW}/\partial u$ has eigenvalues that are all nonnegative and the matrix $\partial g^{SW}/\partial v$ has eigenvalues that are all nonpositive. The resulting expressions for the flux splitting are somewhat simpler when compared to the Steger – Warming splitting. The van Leer splitting also introduces excessive diffusion in the resolution of linearly degenerate contact waves and velocity slip surfaces.

- **System Lax–Friedrichs flux.** This numerical flux is the system equation counterpart of the scalar local Lax–Friedrichs flux (27). For systems of conservation laws, the Lax–Friedrichs flux is given by

$$g^{LF}(u, v) = \frac{1}{2}(f(u) + f(v)) \cdot v - \frac{1}{2}\alpha(v)(v - u)$$

where $\alpha(v)$ is given through the eigenvalues $\lambda_k(v; w)$ of $A(v; w)$

$$\alpha(v) = \max_{1 \leq k \leq m} \sup_{w \in [u, v]} |\lambda_k(v; w)|$$

The system Lax–Friedrichs flux is usually not applied on the boundary of domains since it generally requires an overspecification of boundary data. The system Lax–Friedrichs flux introduces a relatively large amount of artificial diffusion when used in the schemes of Section 2. Consequently, this numerical flux is typically only used together with relatively high-order reconstruction schemes where the detrimental effects of excessive artificial diffusion are mitigated.

- **Harten–Lax–van Leer flux (Harten, Lax and van Leer, 1983).** The Harten–Lax–van Leer numerical flux originates from a simplified two wave model of more general m wave systems such that waves associated with the smallest and largest characteristic speeds of the m wave system are always accurately represented in the two-wave model. The following

numerical flux results from this simplified two-wave model

$$g^{HLL}(u, v) = \frac{1}{2}(f(u) + f(v)) \cdot v - \frac{1}{2} \frac{\alpha_{\max} + \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}} (f(v) - f(u)) \cdot v + \frac{\alpha_{\max} \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}} (v - u)$$

where

$$\alpha_{\max}(v) = \max_{1 \leq k \leq m} (0, \sup_{w \in [u, v]} \lambda_k(v; w))$$

$$\alpha_{\min} = \min_{1 \leq k \leq m} (0, \inf_{w \in [u, v]} \lambda_k(v; w))$$

Using this flux, full upwinding is obtained for supersonic flow. Modifications of this flux are suggested in Einfeldt *et al.* (1998) to improve the resolution of intermediate waves as well.

Further examples of numerical fluxes (among others) include the kinetic flux vector splitting due to Deshpande (1986), the advection upstream spitting method (AUSM) flux of Liou and Steffen (1993), and the convective upwind and split pressure (CUSP) flux of Jameson (1993) and Tatsumi, Martinelli and Jameson (1994).

5 CONCLUDING REMARKS

The literature associated with the foundation and analysis of the finite volume methods is extensive. This article gives a very brief overview of finite volume methods with particular emphasis on theoretical results that have significantly impacted the design of finite volume methods in everyday use at the time of this writing. More extensive presentations and references on various topics in this article can be found in the books by Godlewski and Raviart (1991), Kröner (1997), Eymard, Galluot and Herbin (2000) and LeVeque (2002).

6 RELATED CHAPTERS

(See also Chapter 4, this Volume, Chapter 4, Chapter 11 of Volume 3)

REFERENCES

Abgrall R. On essentially non-oscillatory schemes on unstructured meshes: analysis and implementation. *J. Comput. Phys.* 1994; 114:45–58.

Bank R and Rose DJ. Some error estimates for the box method. *SIAM J. Numer. Anal.* 1987; 24:777–787.

Barth TJ and Frederickson PO. *Higher Order Solution of the Euler Equations on Unstructured Grids Using Quadratic Reconstruction*. Report AIAA-90-0013, American Institute for Aeronautics and Astronautics, 1990.

Barth TJ and Jespersen DC. *The Design and Application of Upwind Schemes on Unstructured Meshes*. Report 89-0366, American Institute for Aeronautics and Astronautics, 1989; 1–12.

Batten P, Lambert C and Causon DM. Positively Conservative high-resolution convection schemes for unstructured elements. *Int. J. Numer. Methods Eng.* 1996; 39:1821–1838.

Billey V, Péraux J, Pernier P and Stoufflet B. *2-D and 3-D Euler Computations with Finite Element Methods in Aerodynamics*, Lecture Notes in Mathematics, vol. 1270. Springer-Verlag: Berlin, 1987.

Boris JP and Book DL. Flux corrected transport: SHASTA, a fluid transport algorithm that works. *J. Comput. Phys.* 1973; 11:38–69.

Bouchut F and Perthame B. Kuzkov's estimates for scalar conservation laws revisited. *Trans. Am. Math. Soc.* 1998; 350(7):2847–2870.

Carrillo J. Entropy solutions for nonlinear degenerate problems. *Arch. Ration. Mech. Anal.* 1999; 147:269–361.

Cai Z. On the finite volume element method. *Numer. Math.* 1991; 58:713–735.

Chatzipantelidis P. A finite volume method based on the Crouzeix–Raviart element for elliptic problems. *Numer. Math.* 1999; 82:409–432.

Chou SH and Li Q. Error estimates in L^2 , H^1 and L^∞ in covolume methods for elliptic and parabolic problems: a unified approach. *Math. Comput.* 2000; 69:103–120.

Cockburn B. Continuous dependence and error estimates for viscosity methods. *Acta Numer.* 2003; 12:127–180.

Cockburn B and Gau H. A posteriori error estimates for general numerical methods for scalar conservation laws. *Comput. Appl. Math.* 1995; 14:37–47.

Cockburn B and Gresham P-A. A Priori Error Estimates for Numerical Methods for Scalar Conservation Laws. Part 1: The General Approach. *Math. Comput.* 1996a; 65:533–573.

Cockburn B and Gresham P-A. A Priori Error Estimates for Numerical Methods for Scalar Conservation Laws. Part 2: Flux Splitting Monotone Schemes on Irregular Cartesian Grids. *Math. Comput.* 1996b; 66:547–572.

Cockburn B and Gresham P-A. A Priori Error Estimates for Numerical Methods for Scalar Conservation Laws. Part 3: Multidimensional Flux-Splitting Monotone Schemes on Non-Cartesian Grids. *SIAM J. Numer. Anal.* 1998; 35:1775–1803.

Cockburn B and Gripenberg G. Continuous dependence on the nonlinearities of solutions of degenerate parabolic equations. *J. Diff. Equations* 1999; 151(2):231–251.

Cockburn B and Shu CW. TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework. *Math. Comput.* 1989; 52:411–435.

Cockburn B and Shu CW. The Runge–Kutta discontinuous Galerkin method for conservation laws. V. Multidimensional systems. *J. Comput. Phys.* 1998; 141(2):199–224.

Cockburn B, Coquel F and Lefloch PG. An error estimate for finite volume methods for multidimensional conservation laws. *Math. Comput.* 1994; 63:77–103.

Cockburn B, Hou S and Shu CW. The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case. *Math. Comput.* 1990; 54(190):545–581.

Cockburn B, Lin SY and Shu CW. TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws. III. One-dimensional systems. *J. Comput. Phys.* 1989; 84(1):90–113.

Cotilla P and Woodward P. The piecewise parabolic methods for gas-dynamical simulations. *J. Comput. Phys.* 1984; 54:174–201.

Chainais-Hillairet C. Finite volume schemes for a nonlinear hyperbolic equation: convergence towards the entropy solution and error estimates. *M2AN Math. Modell. Numer. Anal.* 1999; 33:129–156.

Chainais-Hillairet C. Second-order finite-volume schemes for a non-linear hyperbolic equation: error estimates. *Math. Methods Appl. Sci.* 2000; 23(5):467–490.

Courbade P-H and Debiez C and Dervieux A. *A Positive MUSCL Scheme for Triangulations*. Report 3465, Institut National De Recherche En Informatique Et En Automatique (INRIA), 1998.

Crandall M and Majda A. Monotone Difference Approximations of Scalar Conservation Laws. *Math. Comput.* 1980; 34:1–21.

DiPerna RJ. Measure-valued solutions to conservation laws. *Arch. Rational Mech. Anal.* 1985; 88(3):223–270.

Delanaye M. *Polynomial Reconstruction Finite Volume Schemes for the Compressible Euler and Navier–Stokes Equations on Unstructured Adaptive Grids*. PhD thesis, University of Liège, Belgium, 1996.

Deshpande SM. *On the Maxwellian Distribution, Symmetric Form, and Entropy Conservation for the Euler Equations*. NASA Report TP-2583, NASA Langley: Hampton, Virginia, 1986.

Desideri JA and Dervieux A. *Compressible Flow Solvers Using Unstructured Grids*, Von Karman Institute Lecture Notes, Von Karman Institute for Fluid Dynamics: Belgium, 1988-05.

Einfeldt B, Munz C, Roe P and Sjögren B. On Godunov-type methods near low densities. *J. Comput. Phys.* 1992; 92:272–295.

Evje S and Karlsen KH. An error estimate for viscous approximate solutions of degenerate parabolic equations. *J. Nonlin. Math. Phys.* 2002; 9(3):262–281.

Ewing RE, Lin T and Lin Y. On the accuracy of the finite volume element method based on piecewise linear polynomials. *SIAM J. Numer. Anal.* 2002; 39(6):1865–1888.

Eymard R, Galluot T and Herbin R. Finite volume methods. *Handbook of Numerical Analysis*, vol. 7. North Holland: Amsterdam, 2000; 713–1020.

Eymard R, Galluot T and Herbin R. Finite volume approximation of elliptic problems and convergence of an approximate gradient. *Appl. Numer. Math.* 2001; 37(1–2):31–53.

- Eymard R, Gallouët T and Herbin R. Error estimates for approximate solutions of a nonlinear convection-diffusion problem. *Adv. Diff. Equations* 2002; 7(4):419–440.
- Eymard R, Gallouët T, Ghilani M and Herbin R. Error estimates for the approximate solution of a nonlinear hyperbolic equation given by finite volume schemes. *IMA J. Numer. Anal.* 1998; 18:563–594.
- Eymard R, Gallouët T, Herbin R and Michel A. Convergence of a finite volume scheme for nonlinear degenerate parabolic equations. *Numer. Math.* 2002; 92(1):41–82.
- Feistauer M, Felcman J, Lukášová-Medvid'ová M and Warnecke G. Error estimates for a combined finite volume-finite element method for nonlinear convection-diffusion problems. *SIAM J. Numer. Anal.* 1999; 36(5):1528–1548.
- Gallouët T, Herbin R and Vignal MH. Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions. *SIAM J. Numer. Anal.* 2000; 37(6):1935–1972.
- Godunov SK. A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Math. Sbornik* 1959; 47:271–290.
- Godlewski E and Raviart P-A. Hyperbolic systems of conservation laws. *Mathématiques & Applications*. Ellipses: Paris, France, 1991.
- Goodman JD and LeVeque RJ. On the accuracy of stable schemes for 2D conservation laws. *Math. Comp.* 1985; 45(171):15–21.
- Gottlieb S and Shu CW. Total variation diminishing Runge-Kutta schemes. *Math. Comput.* 1998; 67(221):73–85.
- Gottlieb S, Shu CW and Tadmor E. Strong stability-preserving high-order time discretization methods. *SIAM Rev.* 2001; 43(1):89–112.
- Harten A. High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.* 1983; 49:357–393.
- Harten A. ENO schemes with subcell resolution. *J. Comput. Phys.* 1989; 83:148–184.
- Harten A and Chakravarthy S. *Multi-Dimensional ENO Schemes for General Geometries*. Report ICASE-91-76, Institute for Computer Applications in Science and Engineering, 1991.
- Harten A, Hyman JM and Lax PD. On finite-difference approximations and entropy conditions for shocks. *Commun. Pure Appl. Math.* 1976; 29:297–322.
- Harten A, Lax PD and van Leer B. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Rev.* 1983; 25:35–61.
- Harten A, Osher S, Engquist B and Chakravarthy S. Some results on uniformly high order accurate essentially non-oscillatory schemes. *Appl. Numer. Math.* 1986; 2:347–377.
- Harten A, Osher S, Engquist B and Chakravarthy S. Uniformly high-order accurate essentially nonoscillatory schemes III. *J. Comput. Phys.* 1987; 71(2):231–303.
- Herbin R and Ohlberger M. A posteriori error estimate for finite volume approximations of convection diffusion problems. *proceedings: Finite volumes for complex applications – problems and perspectives, Porquerolles*, Hermes Science Publications: Paris, 2002; 753–760.
- Hermeline F. A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.* 2000; 160(2):481–499.
- Jaffre J, Johnson C and Szepessy A. Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *Math. Models Methods Appl. Sci.* 1995; 5(3):367–386.
- Jameson A. *Artificial Diffusion, Upwind biasing, Limiters and Their Effect on Accuracy and Convergence in Transonic and Hypersonic Flows*. Report AIAA-93-3359, American Institute for Aeronautics and Astronautics, 1993; 1–28.
- Jameson A and Lax PD. Conditions for the construction of multipoint variation diminishing difference schemes. *Appl. Numer. Math.* 1986; 2(3–5):335–345.
- Jameson A and Lax PD. Corrigendum: Conditions for the construction of multipoint variation diminishing difference schemes. *Appl. Numer. Math.* 1987; 3(3):289.
- Jiang G and Shu CW. Efficient implementation of weighted ENO schemes. *J. Comput. Phys.* 1996; 126:202–228.
- Karlsten KH and Risebro NH. On the Uniqueness and Stability of Entropy Solutions of Nonlinear Degenerate Parabolic Equations with Rough Coefficients. Preprint 143, Department of Mathematics, University of Bergen, 2000.
- Koren B. Upwind schemes for the Navier-Stokes equations. *Proceedings of the Second International Conference on Hyperbolic Problems*, Vieweg: Braunschweig, 1988.
- Kröner D. *Numerical Schemes for Conservation Laws*. Wiley-Teubner: Stuttgart, 1997.
- Kröner D and Ohlberger M. A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multidimensions. *Math. Comput.* 2000; 69:25–39.
- Kröner D, Noelle S and Rokyta M. Convergence of higher order upwind finite volume schemes on unstructured grids for conservation laws in several space dimensions. *Numer. Math.* 1995; 71:527–560.
- Kruzkov SN. First order quasilinear equations in several independent variables. *Math. USSR Sbornik* 1970; 10:217–243.
- Küther M. Error estimates for second order finite volume schemes using a TVD-Runge-Kutta time discretization for a nonlinear scalar hyperbolic conservation law. *East-West J. Numer. Math.* 2000; 8(4):299–322.
- Kuznetsov NN. Accuracy of some approximate methods for computing the weak solutions of a first-order quasi-linear equation. *USSR, Comput. Math. Math. Phys.* 1976; 16(6):159–193.
- Lax PD. *Hyperbolic Systems of Conservation Laws*. SIAM: Philadelphia, 1973.
- Lax PD and Wendroff B. Systems of conservation laws. *Commun. Pure Appl. Math.* 1960; 13:217–237.
- Lazarov RD, Miche ID and Vassilevsky PS. Finite volume methods for convection-diffusion problems. *SIAM J. Numer. Anal.* 1996; 33:31–35.
- LeVeque R. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press: Cambridge, 2002.
- Liou MS and Steffen CJ. A new flux-splitting scheme. *J. Comput. Phys.* 1993; 107:23–39.
- Liu X-D. A maximum principle satisfying modification of triangle based adaptive stencils for the solution of scalar hyperbolic conservation laws. *SIAM J. Numer. Anal.* 1993; 30:701–716.
- Málek J, Nečas J, Rokyta M and Růžička M. *Weak and measure-valued solutions to evolutionary PDEs*. *Applied Mathematics and Mathematical Computation*, vol. 13, Chapman and Hall: London, 1968; 44–177.
- Ohlberger M. A posteriori error estimates for finite volume approximations to singularly perturbed nonlinear convection-diffusion equations. *Numer. Math.* 2001a; 87(4):737–761.
- Ohlberger M. A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations. *M2AN Math. Model. Numer. Anal.* 2001b; 35(2):355–387.
- Oleinik OA. Discontinuous solutions of non-linear differential equations. *Amer. Math. Soc. Transl. (2)* 1963; 26:95–172.
- Osher S and Solomon F. Upwind Difference Schemes for Hyperbolic Systems of Conservation Laws. *Math. Comput.* 1982; 38(158):339–374.
- Osher S. Riemann solvers, the entropy condition, and difference approximations. *SIAM J. Numer. Anal.* 1984; 21(2):217–235.
- Peterson T. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *SIAM J. Numer. Anal.* 1991; 28(1):133–140.
- Roe PL. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.* 1981; 43:357–372.
- Rostand P and Stottfort B. TVD schemes to compute compressible viscous flows on unstructured meshes. *Proceedings of the Second International Conference on Hyperbolic Problems*. Vieweg: Braunschweig, 1988.
- Shu CW. Total-variation-diminishing time discretizations. *SIAM J. Sci. Statist. Comput.* 1988; 9:1073–1084.
- Shu CW. *High Order ENO and WENO Schemes, Lecture Notes in Computational Science and Engineering*, vol. 9, Springer-Verlag: Heidelberg, 1999.
- Shu CW. A survey of strong stability preserving high order time discretizations. *Collected lectures on the preservation of stability under discretization* (Fort Collins, CO, 2001), SIAM: Philadelphia, 2002; 51–65.
- Shu CW and Osher S. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comput. Phys.* 1988; 77:439–471.
- Sonar T. On the construction of essentially non-oscillatory finite volume approximations to hyperbolic conservation laws on general triangulations: polynomial recovery, accuracy, and stencil selection. *Comput. Methods Appl. Mech. Eng.* 1997; 140:157–181.
- Sonar T. On families of pointwise optimal finite volume ENO approximations. *SIAM J. Numer. Anal.* 1998; 35(6):2350–2379.
- Spekreijse SP. Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws. *Math. Comput.* 1987; 49:135–155.
- Steger JL and Warming RF. Flux vector splitting of the inviscid gasdynamic equations with application to finite difference methods. *J. Comput. Phys.* 1981; 40:263–293.
- Sitli E. Convergence of finite volume schemes for Poisson's equation on nonuniform meshes. *SIAM J. Numer. Anal.* 1991; 28:1419–1430.
- Sweby PK. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.* 1984; 21(5):995–1011.
- Szepessy A. An existence result for scalar conservation laws using measure valued solutions. *Commun. Partial Diff. Equations* 1989; 14:1329–1350.
- Tadmor E. Local error estimates for discontinuous solutions of nonlinear hyperbolic equations. *SIAM J. Numer. Anal.* 1991; 28:891–906.
- Tartar L. The compensated compactness method applied to systems of conservation laws. *Systems of Nonlinear Partial Differential Equations*. Reidel: Dordrecht, 1983.
- Tatsumi S, Martinelli L and Jameson A. *Design, Implementation, and Validation of Flux Limited Schemes for the Solution of the Compressible Navier-Stokes Equations*. Report AIAA-94-0647, American Institute for Aeronautics and Astronautics, 1994; 1–20.
- Vankeirsbliek P. *Algorithmic Developments for the Solution of Hyperbolic Conservation Laws on Adaptive Unstructured Grids*. PhD thesis, Katholieke Universiteit Leuven, Belgium, 1993.
- Van Leer B. Towards the ultimate conservative difference schemes V: A second order sequel to Godunov's method. *J. Comput. Phys.* 1979; 32:101–136.
- Van Leer B. *Flux-Vector Splitting for the Euler Equations*. Report ICASE-82-30, Institute for Computer Applications in Science and Engineering, 1982.
- Van Leer B. *Upwind-Difference Schemes for Aerodynamics Problems Governed by the Euler Equations, Lectures in Applied Mathematics* 22, AMS: Providence, Rhode Island, 1985.
- Vila JP. Convergence and error estimates in finite volume schemes for general multi-dimensional scalar conservation laws I: Explicit monotone schemes. *RAIRO, Model. Math. Anal. Numer.* 1994; 28:267–295.
- Viozat C, Held C, Mer K and Dervieux A. *On Vertex-Center Unstructured Finite-Volume Methods for Stretched Anisotropic Triangulations*. Report 3464, Institut National de Recherche En Informatique Et En Automatique (INRIA), 1998.
- Wierse M. *Higher Order Upwind Scheme on Unstructured Grids for the Compressible Euler Equations in Time Dependent Geometries in 3D*. PhD thesis, University of Freiburg, Germany, 1994.

Chapter 16

Geometric Modeling of Complex Shapes and Engineering Artifacts

F.-E. Wolter, N. Peinecke and M. Reuter

University of Hannover, Hannover, Germany

| | |
|------------------------------------|-----|
| 1 Architecture of Modeling Systems | 475 |
| 2 Voxel Representation | 476 |
| 3 Surface Patches | 477 |
| 4 Boundary Representation | 481 |
| 5 Constructive Solid Geometry | 483 |
| 6 Medial Modeling | 485 |
| 7 Attributes | 490 |
| 8 Outlook and Concluding Remarks | 492 |
| Acknowledgments | 494 |
| Notes | 494 |
| References | 494 |

1 ARCHITECTURE OF MODELING SYSTEMS

It is not easy to define a modeling system. A modeling system can be every system useful to model a 2-D or 3-D object. Still many designers model with clay, hence from their point of view a modeling system would be pencil, paper, clay, and the designer himself. In the area of computer graphics, one is mainly interested in a virtual model of the object, which can be viewed from different perspectives, modified, and processed further, to simulate the behavior of the object in reality. Here a modeling system consists of the computer hard- and software and of the user.

Encyclopedia of Computational Mechanics, Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes. Volume 1: *Fundamentals*. © 2004 John Wiley & Sons, Ltd. ISBN: 0-470-84699-2.

Before one can choose or build a modeling system, one has to choose an appropriate model for the problem given. We will discuss different types of models in the following chapters but we will not go into detail on how to map a given real-world problem onto one of these models. Please refer to Koenderink (1990) for some insights on how to accomplish this.

Today a strict separation of physical modeling, for example, with clay and virtual modeling with a computer, cannot be sustained, since many mixtures are used in practice. Clay modelers, for example, often use a 3-D scanner to create a virtual model and on the other hand virtual models can easily be printed with a 3-D printer to create 3-D prototypes. Recently, even stronger connections are made using haptical devices and 3-D glasses to enable the user to feel and see the object in 3-D space.

Since the user is still the most important part of a modeling system, the interaction between the human and the computer plays a crucial role. Therefore, different hardware tools like scanners, printers, viewing, and input devices have been developed to interact with the user. The software is then needed to ensure the smooth interaction of all components.

The software of a modeling system can be divided into four abstraction layers (see Figure 1):

1. The *user interface* (UI) is the part of the software that interacts with the user directly. The UI is mostly graphical and presents the user with many options to create, modify, analyze, and view the object. Constructing a graphical UI is a complex venture by which not only the wishes of the user have to be taken into consideration, but also the possibilities of the hardware. It is important that operations being repeated very often do not consume too much time and that the user is

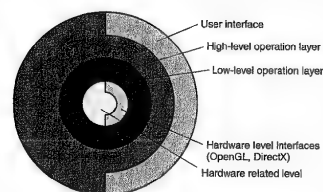


Figure 1. Software levels of a modeling system. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

constantly informed about the status of any operation. An intuitive layout (of buttons, menus...) should also be kept in mind.

2. The *high-level operation* layer hosts mainly complex operations like intersecting, cutting, modifying, analyzing, and postprocessing of objects. These operations can be accessed through the user interface and can be understood to be the main modeling tools. They should be robust and efficient to supply the user with powerful tools enabling him to achieve every option he has in mind.
3. On the *low-level operation* layer, the data structure is located together with its low-level operators. These operators provide the next higher level with the controlled access and modifying options of the data structure. They keep the data in an organized state. Since the data structure and its operators are strongly connected, an object oriented programming language like C++ is predestined for the implementation.
4. The *hardware-related level* is the lowest layer. Here the interaction with the input- and output-hardware devices is implemented. Sometimes it is necessary to directly program the hardware (driver programming, assembly language, etc.) to elicit the needed features, but most of the time it is sufficient to use existing drivers and interfaces (e.g. OpenGL or DirectX). Another important aspect that needs to be dealt with on the lowest layer is the precision of operations. Since floating-point arithmetic is only approximate, but not precise, small errors may accumulate and possibly lead to catastrophic failure. Therefore provisions have to be made to prevent this failure or an exact arithmetic has to be implemented, unfortunately leading to a slowdown of the entire system.

We have seen that the data structure and its operators form the heart of the modeling system (level 3).

Therefore the data structure determines the feasibility and performance of the high-level operations. Many different types of data structures exist, each with its own advantages (and disadvantages).

More on modeling systems (with an approach slightly different from the one presented here) can be found in Hoffmann (1989).

2 VOXEL REPRESENTATION

A typical volume-based approach in modeling is the *voxel representation*. Koenderink (1990) refers to these kinds of models as "sugar cube blobs" since these models can be thought as a set of sugar cubes glued together appropriately. This concept is a straightforward generalization of pixel graphics as known from computer graphics. Whereas in pixel representations a 2-D image is discretized into a set of squares with integer coordinates (the pixels), in voxel representations 3-D space is split into a regular cubic grid consisting of voxels. The easiest way of representing an object like this is to assign to each voxel a Boolean value, deciding if the volume described by the voxel is part of the object or not. Figure 2 shows a typical $12 \times 12 \times 12$ representation of a full sphere. A problem of the voxel-based approach is that the approximation of the objects volume at low resolutions is usually relatively poor, while higher resolutions increase memory consumption at a cubic rate. As a compromise for voxels intersecting the boundary of the object, the Boolean values can be changed to fuzzy numbers depending on the volume of the intersecting part of voxel and object. Additional attributes as described later can also be assigned to voxels.

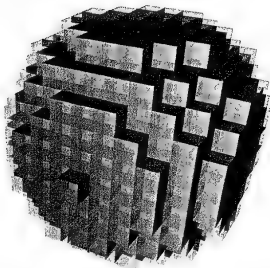


Figure 2. Voxel representation of a full sphere. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

Voxel representations make Boolean operations like intersection or union of two objects extremely easy. Only the corresponding Boolean operations for their assigned voxel values need to be carried out, for example the logical AND for the intersection. Again this is relatively costly at a higher resolution due to the enormous number of voxels involved.

The voxel-representation method can be viewed as a special case of the constructive solid geometry (CSG) technique discussed in Section 5 with only one primitive – the cube at integer coordinates – and one operator – the union.

Voxel representation is also known as *spatial-occupancy enumeration* (cf. Foley et al., 1996).

2.1 Octrees

Voxel representations can become memory consuming if a greater level of detail is desired. Thus sometimes voxels are organized into octrees. These are trees where each node is of degree eight or zero. Octrees are obtained by starting with one voxel large enough to enclose the whole object, representing the root node of the octree. In general this voxel is a poor approximation of the object, therefore this voxel is divided into eight equal-sized smaller voxels, representing the child nodes of the root node. For each voxel, an approximation criterion is checked, for example if the voxel intersects the boundary of the object. If this criterion is met, it is subdivided further, otherwise subdivision is omitted. This process is repeated for those voxels that require further subdivision until the desired level of approximation is reached.

To understand the way an octree is obtained refer to Figure 3. Here the octree's 2-D analogue, a quadtree, for a triangle is constructed. The resulting quadtree is shown in Figure 4. Note that only squares (and thus nodes) contributing to a greater level of detail are to be refined in a following step. Hierarchical representation schemes like octrees make tasks like collision detection particularly easy: First the two root nodes need to be checked for intersection. If and only if an intersection is found, the child nodes belonging to the respective objects are checked and so on. This is almost as easy as in the voxel case while being far more efficient. For an overview on how to implement Boolean operations for octrees, see Foley et al. (1996). For



Figure 3. A quadtree for a triangle. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

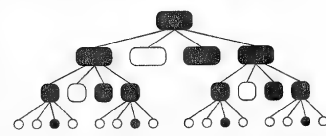


Figure 4. Resulting quadtree for the triangle. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

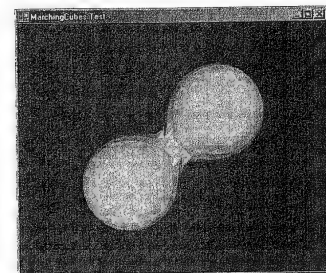


Figure 5. Result of a marching cubes conversion. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

a comprehensive survey of octree-related techniques, see Samet (1984).

Both voxel and octree representations may require conversion to a boundary-representation before FEM computations can be carried out. This conversion can be accomplished using the famous *marching cubes algorithm* (Lorensen and Cline, 1987). Figure 5 depicts the result of such a conversion.

3 SURFACE PATCHES

Surface patches form the base for boundary-representation schemes. Therefore before we can discuss the foundations of boundary representations in Section 4, we need to know how to model a surface – the boundary of our object.

We define a *surface patch* to be a connected 2-D manifold in 3-D, that is, a set of points in 3-D, where each inner point has a small surrounding neighborhood homeomorphic to the

2-D open disc. We define the boundary of this set to be part of the surface patch.

There exists quite a huge variety of surface patches matching this definition and most of them are not easily represented in a data structure. Furthermore, we should keep in mind that surface patches are usually meant to be 'sewn' together (cf. Section 4) in order to form more complex surfaces, therefore they are mainly rather simple bounded and bordered manifolds. Nevertheless, we shall give no formal definition of simplicity but rather present a selection of commonly used techniques for implementing special classes of surface patches.

For a detailed discussion of many of the subjects mentioned in this section refer to Hoschek and Lasser (1993). Also refer to Koenderink (1990) for some deeper insights about patches.

3.1 Polygonal patches

Given a sequence of coplanar points p_0, \dots, p_n in 3-D, we define the sequence of edges joining two points $\overline{p_i, p_{i+1}}$ plus the edge $\overline{p_n, p_0}$ to be the *closed polygon* of the points. We define the *geometric interior* of the polygon to be the set of all points in the same plane that cannot be reached by an arbitrary path from a point far away in the same plane (e.g. from a point outside the convex hull of the polygon) without crossing the polygon. We will consider every geometric interior point plus the boundary of these points to be part of the polygonal patch.

Of course this definition gives no efficient algorithm for testing, if a point belongs to the polygonal patch. There exists a variety of methods to do this (cf. Foley et al., 1996), each meeting our definition in special cases (and not in others). For example, some efficient algorithms fail if the polygon is not simple, that is, possesses self intersections. Figure 6 shows different polygons with their geometric interior painted red. It is also often desirable to allow polygons to have inner-boundary components, that is, inner parts of the boundary that are not directly connected to the outer boundary. We will refer to these as *inner loops* (cf. Section 4). Figure 7 shows a rectangular polygon with two inner loops. Note that these polygons also match our definition of a surface patch.



Figure 6. Geometric interior of polygons. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

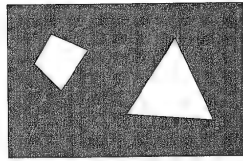


Figure 7. Polygon with inner loops. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

Because every polygonal patch can be decomposed into a set of triangular patches, it is sometimes sufficient to consider only triangular patches, yielding its *triangulation*. These can be handled very efficiently especially inside-outside testing and various other calculations are easily carried out for triangles. Nevertheless since a triangulation is not unique for a patch, a chosen triangulation sometimes introduces biases into these calculations. Often elaborate meshing techniques yielding almost equilateral triangles need to be applied. Therefore more general schemes allowing also nontriangular patches should be carefully considered as well. Triangulation is a special case of *meshing techniques*. Figure 8 shows an example of an object composed of polygonal patches (here only triangles).

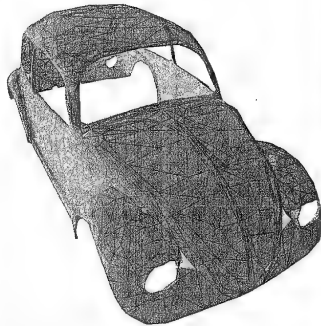


Figure 8. Triangulated object. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

3.2 Parametric surfaces

Often we want the surface to be really curved instead of just (piecewise) planar like in the polygonal case. This can be achieved via *parametric surfaces*. Let D be a subdomain of 2-D space and let $f: D \rightarrow \mathbb{R}^3$ be a continuous map. Often we require f to be differentiable, mostly f will be a homeomorphism onto its image set $f[D]$ and generally we assume the differential of f having maximal rank. We will call the pair (D, f) a *parametric surface* with parameterization f , the surface patch is represented by the image of f [1].

Note that we assume no further restrictions for D , allowing explicitly every planar polygon (in 2-D), even with inner loops. This is because it can be sometimes intricate to find parameterizations for special surfaces. For example, a ring-like structure as depicted in Figure 9 can easily be represented by the domain

$$[-1, 1] \times [-1, 1] \setminus \left[-\frac{1}{3}, \frac{1}{3} \right] \times \left[-\frac{1}{3}, \frac{1}{3} \right]$$

with parameterization

$$f(x, y) := \frac{\max\{|x|, |y|\}}{\sqrt{x^2 + y^2}} \begin{pmatrix} x \\ y \\ 0 \end{pmatrix}$$

Nevertheless, most of the time D will be polygonal or – for convenience – the unit square. The spline surfaces discussed in Section 3.3 are a popular example of parametric surface patches. An alternative method is to model using partial differential equations. This elaborate technique was developed by M. Bloor and M. Wilson at the University of Leeds and is described in Nowacki, Bloor and Oleksiewicz (1995).

3.3 Spline surfaces

It is well known that for a given set of 3-D points p_0, \dots, p_n there is a unique polynomial curve

$$\alpha(t) = \sum_{i=0}^n c_i t^i$$

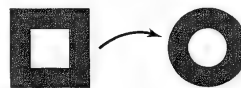


Figure 9. Parameterization of a planar ring. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

with $c_0, \dots, c_n \in \mathbb{R}^3$ such that α interpolates every point p_j . We refer to the points c_i as *control points* of α . Polynomial interpolants suffer from three major drawbacks:

- They tend to form unexpected "swinging" curves that can move far away from the interpolation points.
- Construction and evaluation of these curves are numerically unstable.
- There is no intuitive interrelation between coefficients of a polynomial and the shape of the resulting curve or surface.

Splines try to overcome all three problems by two basic techniques:

- Use a type of curve, that is numerically and visually more "tame", that is, closer to its interpolation points.
- Compose the curve piecewise from subcurves.

All different types of splines are obtained from piecewise subcurves, they only differ by the base type chosen for these curves. Best known and widely used are Hermite-splines, Bezier-splines, B-splines and NURBS, and of course monomial-splines (where each subcurve is an ordinary polynomial curve). These are all curves of piecewise polynomial type (and in the case of NURBS piecewise rational polynomial). Furthermore there are nonpolynomial types like trigonometric splines, exponential splines, or splines based on subcurves obtained from other subdivision processes (which are not necessarily polynomial), although these are more rarely used as they may be computational costly.

Formally we will call a curve a *spline* (of degree n) if it is

1. piecewise composed of the same type of subcurve belonging to the same finite dimensional vector space of functions (note that the subcurve is in most cases C^∞),
2. at least $n-1$ times continuously differentiable.

Note that depending on the type of spline chosen we often need additional control points besides the interpolation points to characterize the curve completely.

Using the techniques from Section 3.2 one can easily obtain *spline surface patches* from spline curves. Given an array of control points (c_{ij}) with $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$, each sequence c_{1j}, \dots, c_{nj} defines a spline curve α_j , which evaluated at a certain point x yields a further sequence of control points $\alpha_1(x), \dots, \alpha_n(x)$. These form a spline β_x , which can then be evaluated at a point y , thus giving a resulting range point. This process describes

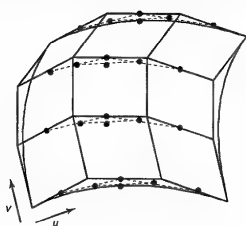


Figure 10. Control array of a spline surface.

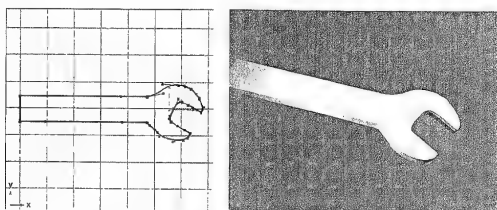
a map

$$f(x, y) := \beta_x(y)$$

where f in fact is a parameterization of a surface patch. Figure 10 shows a control array and the underlying spline surface.

For a detailed overview on spline techniques see, for example, Farin (1993), Hoschek and Lasser (1993), and Yamaguchi (1988). A recent reference can be found in Patrikalakis and Maekawa (2002), this book also deals with problems of spline surface intersections, which are important when splines are combined with CSG representations (cf. Section 5). Certainly the most elaborate spline technique is the usage of *NURBS* (nonuniform rational B-splines); see Piegl and Tiller (1995) for a comprehensive overview.

Figure 11 shows a wrench modeled with piecewise B-spline patches.

Figure 11. Wrench composed of B-spline patches. A color version of this image is available at <http://www.mrw.interscience.wiley.com/cem>

3.4 Trimmed surfaces

We have already seen in Section 3.2 that the domain of a parametric surface is not necessarily the unit square. We can generalize this principle by trimming polygonal and even nonpolygonal (e.g. bounded by splines) subdomains from parameter space and thus trimming the surface patch itself. Depicted in Figure 12 you find a sequence in which a user selects a closed curve in parameter space (green), which is then trimmed out of the surface patch (black). Note that trimming the surface patch directly (instead of the parameter domain) is an even more complex task since it involves the computation of the inverse f^{-1} of the parameterization f (c.f. Patrikalakis and Maekawa, 2002).

3.5 Multiresolutional approaches

As we have seen in Section 3.3, splines bring great improvements in curve and surface design. Nevertheless, modern design and modeling applications may demand further features of a surface representation that are in detail (cf. Stollnitz, DeRose and Salesin, 1996):

- easy approximation and smoothing of a representation, gained from external sources (i.e. scan points from a digitizer)
- changing the macro-scale appearance without changing the micro-scale appearance (the "character") and vice versa
- edit a representation on various, preferably continuous, levels of detail

These issues can be addressed by using a relatively new idea, the so called *B-spline wavelets* instead of ordinary B-splines for curve and surface modeling.

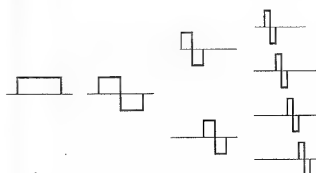
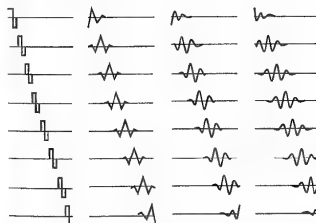
Figure 12. Trimming of a surface patch. A color version of this image is available at <http://www.mrw.interscience.wiley.com/cem>

Figure 13. Haar base wavelets.

Figure 14. Some B-spline wavelets. (Reprinted from *Wavelets for Computer Graphics: Theory and Applications*, Stollnitz E.J., DeRose T.D. and Salesin D.H., *The Theory of Multiresolutional Analysis: Biorthogonal Wavelets*, (1996), p. 96, with permission from Elsevier.)

The idea of a wavelet representation is to represent a curve using linear combinations of functions given in different levels of detail. The *Haar* base functions shown in Figure 13 are the best known examples. Thus we get a representation in which a manipulation of a single coefficient has only relatively local impact, depending on the level. Since B-splines also form a vector space, wavelets can be

built from them. Figure 14 depicts B-spline wavelets for degree 0 up to degree 3 for the third level of detail (thus giving $2^3 = 8$ functions per degree). For a comprehensive overview on wavelets from an analytical point of view, refer to the book by Mallat (1998).

4 BOUNDARY REPRESENTATION

A boundary representation (B-rep) of a solid describes the solid only by its oriented boundary surface. The orientation is needed to decide easily which side of the boundary is the top side and which is the bottom side (even if the object is not closed). Since a normal vector is known everywhere B-rep solids can be visualized very easily.

Generally it is possible to use a variety of different surface patches to model the boundary. These patches (e.g. NURBS-patches, parameterized surfaces, or simply planar polygons, see Section 3) have to be connected with each other at their boundaries. The orientation must not be destroyed during this step.

In most applications, planar polygons are used as patches (very often only triangles are permitted). These patches are called *faces*. Their borders consist of *vertices* and *edges*. Different data structures have been developed to hold the information necessary to create and work with a B-rep solid. The location and orientation (normal vector) of a plane containing a face has to be known and also the correspondence of the vertices and adjacencies of the edges and faces need to be controlled.

The boundary representation of a solid therefore has two parts:

- the *topological* description of the connectivity and orientation and
- the *geometric* description to place all elements in space.

To understand how the topological data are maintained and by what means a topological integrity can be ensured, it is useful to understand *planar models* and *Euler operators* first. Later on a *half-edge data structure* is

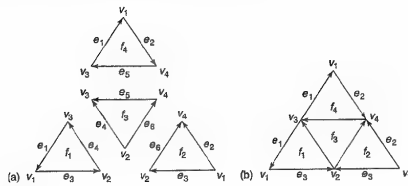


Figure 15. Planar model of a tetrahedron.

introduced as an example on how to implement a B-rep model (see Chapter 7, this Volume).

4.1 Planar models

Planar models are useful to represent the topology of a solid. A planar model is a planar oriented graph (F, E, V) consisting of a set of faces $F = \{f_1, f_2, \dots\}$, edges $E = \{e_1, e_2, \dots\}$, and vertices $V = \{v_1, v_2, \dots\}$. Every edge has its orientation. If different faces share the same edges or vertices, they have to be identified with each other. Identified edges must show in the same direction. In other words, the directions of the edges imply the way they have to be identified. Note that with the half-edge data structure described in the next chapter, one edge always consists of two half-edges showing in opposite directions. Here we only have a single edge, which can appear several times in a planar model. Figure 15 shows an example of the planar model of a tetrahedron. Figure 16 shows a model of the torus and the Klein bottle. The only difference between the two models in Figure 16 is that one of the edges e_2 points in the opposite direction. This results in a different identification. The two models therefore describe different objects.

A solid can have different planar models. An example can be found in Figure 17, where two different planar models of the sphere are presented.

From a planar model, the Euler characteristic can be calculated quickly: $\chi = |V| - |E| + |F|$. Here it does not matter which particular planar model of a solid is used. The Euler characteristic of the tetrahedron is $\chi_T = 4 - 6 + 4 = 2$, the characteristic of the torus and Klein bottle (Figure 16) is $\chi_B = 1 - 2 + 1 = 0$, and of the sphere (Figure 17) is $\chi_S = 1 - 1 + 2 = 2 - 2 + 2 = 2$.

Based on the planar models, a set of topological operators was developed to manipulate models in a way that leads to all models of physical significance while making sure that only feasible models can be created. For instance,

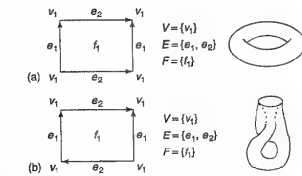


Figure 16. Planar model of the (a) torus (b) Klein bottle.

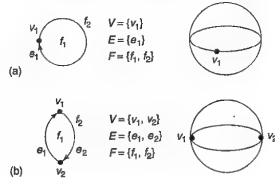


Figure 17. Planar models of the sphere.

nonorientable models cannot be created. These topological operators are split into two classes, the local and global operators.

Local operators work on planar models without modifying the Euler characteristic while global operators can create objects of higher genus (e.g. double torus), thus changing the Euler characteristic. A detailed description and proofs of the properties of these operators can be found in Mäntylä (1988).

4.2 Half-edge data structure

In order to work with a B-rep model, one needs to construct a data structure, combining geometric and topological data. The probably oldest formalized structure, the so called *winged-edge* data structure, was introduced by Baumgart (1975). The half-edge data structure is a variation by Mäntylä (1988) that permits multiple connected faces and sustains a close relationship to the planar models.

The half-edge data structure (as depicted in Figure 18) utilizes the fact that each edge of the boundary surface of a closed solid belongs to exactly two faces. So every edge is split into two half-edges that are oriented in opposite directions. Every face has exactly one outer boundary (outer loop) consisting of counterclockwise oriented half-edges (if viewed from above) and possibly further inner loops consisting of half-edges that are oriented clockwise. The orientation of the loops makes it possible to determine the top and the bottom side of each face. All vertices of a face have to lie on the same plane and are saved in 3-D homogeneous coordinates. Every vertex has to be unique and can be referenced by many half-edges (depending on how many faces share that vertex). Since all half-edges know their neighbor and their parent loop, which again knows the parent face, finding neighbour faces and iterating through the data structure is quite easy.

A set of low- and high-level operators (the so called *Euler operators*) can be derived from the topological operators of the planar model (see the previous section). This permits operations on the data structure in an ordered manner. Any further operators can be implemented using the Euler operators, thus granting the technical feasibility of the modeled object (see Chapter 17 and Chapter 18 of this Volume).

5 CONSTRUCTIVE SOLID GEOMETRY

One of the best known volume-based approaches to modeling is the CSG approach. Again refer to Foley *et al.* (1996) or Hoffmann (1989) for an overview on the subject.

In CSG, every object is either one of a set of simple objects, the *primitives* or it is derived from these by a sequence of operations. Various CSG schemes exist. They are different with respect to their sets of primitives and operations. In 3-D modeling, the most commonly used primitives are:

- ball,
- cylinder,
- box,
- cone.

Further possibilities include surfaces of revolution, implicit bodies, and boundary-representation objects. This shows

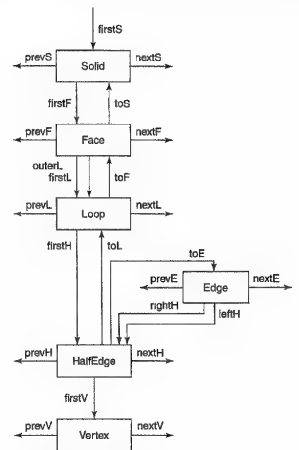


Figure 18. Half-edge data structure.

that CSG can be combined with other methods (discussed above) to gain a greater variety of primitives.

A suitable set of operations must include:

- Euclidean motions (translation, rotation)
- union,
- intersection,
- difference.

The latter three are called *regularized Boolean operations* because they are analogous to the well-known Boolean set operations with a slight difference we will discuss later. Let us first consider the example shown in Figure 19. The object on the left side is composed of the primitives on the right side via the union operation. Note that parts of the objects located inside other objects are 'swallowed', so that there are no more overlaps or double points.

Internally composite objects are kept as binary operator trees. Figure 20 shows one of such trees, U denotes the union operator. Obviously neither the sequence of operators nor the resulting tree is unique for a given result. Nevertheless, by this way, CSG keeps a kind of history of the construction steps, hence every complex object can be



Figure 19. An object composed of CSG primitives. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

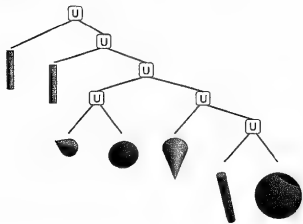


Figure 20. CSG tree for the bird object. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

decomposed into primitive parts again. This is not possible with most other methods.

Figure 21 shows another example, a wrench composed of the primitives shown on the right. Formally the regularized Boolean operators are defined as follows: Given two objects A and B and a Boolean set operator \circ . The result of the corresponding regularized Boolean operator \circ is defined to be $A \circ B := \bar{A} \circ \bar{B}$, where A^* denotes the interior of A and \bar{A} denotes the closure. This definition avoids problems of Boolean operators giving results that do not represent 3-D objects. For example, the intersection of two adjacent boxes sharing one side would yield just that side as a result, giving a non 3-D object.

Sometimes further operations are included like non-Euclidean matrix operations (scaling, skew transforms) or surface sweep operations. One has to take care that these additional operations are applicable for the given primitives, for example, it is mostly impossible to apply sweep operations to objects given in implicit representations while keeping their implicit representation.

CSG is best suited for ray tracing and other applications that require only inside-outside testing as these can be

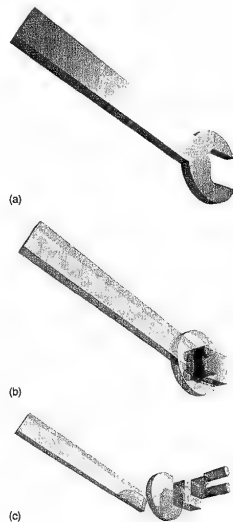


Figure 21. A CSG wrench model. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

easily carried out in a CSG representation. Also voxel representations can be gained easily from a CSG representation. On the other hand, CSG is not easy to deploy in situations that require a meshing of the given object, for example,

for FEM computations since this demands the elimination of unnecessary (e.g. 'swallowed') object parts first which can be costly. A method that avoids these computations is to apply the marching cubes algorithm (Lorensen and Cline, 1987). This only requires that it can be checked for an arbitrary point if this point is inside the interior of an object, which is easily possible for a CSG representation. One major drawback here is that we might lose important details of the model. Another method would be to mesh each primitive separately (most of the time, it is relatively easy to give a mesh for each primitive) and join these (this is the difficult part) to form the mesh.

6 MEDIAL MODELING

Medial modeling is the newest one among the modeling concepts presented in this survey paper. Medial modeling essentially uses past and ongoing research of the Welfenlab being the authors research lab at the University of Hannover (cf. Wolter and Frieze (2000) for a brief overview of results). The suggestion to use the medial axis concept as a tool for shape interrogation appears to have been discussed first quite extensively by Blum (1973). For a detailed mathematical analysis of the underlying mathematical concepts (in the generalized context as cut loci), refer to Wolter (1979, 1985, 1992). The latter paper presenting an extended analysis of mathematical foundations of the medial axis contains also early results (e.g. the one stated below in formula 1) indicating already the possibility to employ the medial axis as a geometric modeling tool. The latter aspect will be a subject discussed in this section.

One could perhaps summarize the most relevant points of medial modeling as follows: In medial modeling, a solid is described by its medial axis and an associated radius function. The medial axis being a subset of the solid is a collection of lower dimensional objects. For a general 3-D solid, the medial axis mostly consists of a connected set built by a collection of surface patches and curves. Medial representations often simplify the process of gaining volume tessellations for the given object, supporting the meshing of solids, cf. Section 6.4. They also offer new possibilities for the construction of intuitive haptic user interfaces that are useful to mold a solid's shape.

The basis of medial modeling can be summarized in a few geometric observations, ideas, and definitions that are outlined here. Let K be a solid in the 3-D or 2-D Euclidean space. The medial axis $M(K)$ of K being a subset of the solid contains all points in K that are centers of maximal discs included in K . One usually includes in the medial axis set $M(K)$ its limit points. Figure 22 shows the medial axis (green) of a domain (black) with some maximal discs

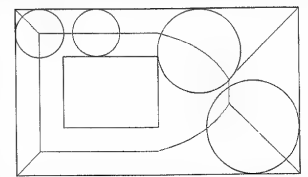


Figure 22. Medial axis of a domain. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

given (red). We assign to any point p in the medial axis $M(K)$ the radius $r(p)$ of the aforementioned maximal disc with center p and radius $r(p)$. This disc is denoted with $B_{r(p)}(p)$. The pair $(M(K), r)$ described by the medial axis $M(K)$ of a solid K and the associated maximal disc radius function

$$r: M(K) \rightarrow \mathbb{R}^+$$

is called *medial axis transform*, where \mathbb{R}^+ denotes the nonnegative real numbers. This pair $(M(K), r)$ yields a new possibility to represent the solid K simply as the union of the related maximal discs, that is,

$$K = \bigcup_{p \in M(K)} B_{r(p)}(p) \quad (1)$$

For details see Wolter (1992). Figure 23 shows how the union of maximal discs defines the shape of a planar domain. The general reconstruction statement expressed in equation 1 already holds for solids with merely continuous boundary surfaces. However, if the solid has merely continuous boundary surfaces (or continuous boundary curves for solids being closed 2-D domains) then the medial axis may have a "wild" structure that may, for example, be presented by a set being dense in some open 3-D sets (containing 3-D discs).

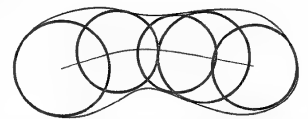


Figure 23. Maximal discs defining a shape. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

In case one poses some regularity conditions for the solid's boundary surface ∂K , for example, curvature continuity, then the respective medial axis $M(K)$ will have a more benign structure. For instance if ∂K is curvature continuous, then $M(K)$ will not be dense in any open set in \mathbb{R}^3 .

Let us assume that ∂K is built by a finite collection of finitely many B-spline patches, then $M(K)$ could be constructed by a union of finitely many medial sets. Each of the latter consisting of points being equidistant to two appropriately chosen boundary parts. Hence each medial set can be viewed as subset of zero sets defined implicitly by the condition stating that the difference of the distances of a point in the medial set to the respective parts of ∂K is zero.

This insight can be used to develop strategies to compute (approximately) the medial axis by assembling it from medial sets (see Figure 24). In case the boundary parts are given implicitly by solutions to polynomial equations, then the medial sets can be described in principal by implicit polynomial equations as well.

The reconstruction result stated above in equation 1 can be used also to model shape for families of objects. Clearly, in equation 1, the shape of the object depends on the medial axis set $M(K)$ and the function r .

Intuitively a continuous deformation $M(K)_t$, $t \in \mathbb{R}_+$, of the medial axis $M(K) = M(K)_0$ combined with a continuous change of the function r described via a continuous family of functions $r(t, s) : M(K) \rightarrow \mathbb{R}^+$ with $r(0, 0) : M(K) \rightarrow \mathbb{R}$ (with $r(0, 0) = r$) should yield a continuously deformed family of objects

$$K_{(t,s)} = \bigcup_{p \in M(K)} B_{r(t,s)(p)}(p)$$

The two control parameters t, s indicate that the change of the radius function $r(t, s)$ depends on the chosen respective domain of definition controlled by the parameter t and

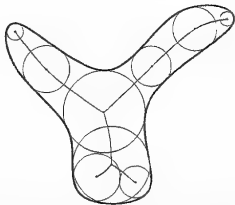


Figure 24. Assembled shape. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>



Figure 25. Continuous deformation of an object. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

for a fixed domain of definition $M(K)$, the radius function depends on the parameter s . Figure 25 shows such a deformation. Note that the medial axis and the radius function are both modified. In order to present a well-defined concept for the continuity of the deformation outlined here, we need some formal requirements that are caused by some complications that may occur during the deformation process. We observe that a continuous (differentiable) deformation of the solid's boundary may result in a family of medial axes $M(K)$, whose homeomorphic type may change during the deformation process. Such a metamorphosis will occur when a (new singularity) curvature center of the boundary will meet the family of medial axes occurring during the deformation process of the solid. See Figure 26 for an example. Therefore it makes sense to consider continuously changing families of functions $r(s, t)$ under the provision that for a varying parameter s the domain of the function family, here the medial axis set $M(K)_t$, should be fixed. This will allow to consider (for a fixed parameter t_0 and a variable parameter s) the family of continuous functions $r(t_0, s) : M(K)_{t_0} \rightarrow \mathbb{R}$ as a continuous path in a vector space of real-valued functions defined on the compact set $M(K)_{t_0}$. That vector space will be endowed with an appropriate topology or norm. Here, fixing the domain $M(K)$, makes it easy to define a distance between two radius functions $r(t_0, s_1)$ and $r(t_0, s_2)$ by

$$d(r(t_0, s_1), r(t_0, s_2)) := \max_{p \in M(K)_{t_0}} \{|r(t_0, s_1)(p) - r(t_0, s_2)(p)|\}$$

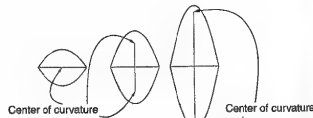


Figure 26. Nonhomeomorphic deformation of a medial axis. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

In order to express the continuous deformation of the medial axis family in a formally precise setting, we need to endow the collections of all medial axes with a topology as well. Here it makes sense to use the Hausdorff metric $d_H(\cdot, \cdot)$ defined on all compact subsets of the respective 2-D or 3-D domain. Let $A, B \subset \mathbb{R}^2$ then

$$d_H(A, B) := \inf\{\epsilon : A \subseteq N_\epsilon(B) \text{ and } B \subseteq N_\epsilon(A)\}$$

with

$$N_\epsilon(A) := \{y : |x - y| < \epsilon \text{ for some } x \in A\}$$

A continuously deformed family of medial axes (depending on a parameter t) can now be viewed as a continuous path ϕ in the Hausdorff space H_{d_H} of compact sets in \mathbb{R}^2 or \mathbb{R}^3 . Here we have

$$\phi(t) : \mathbb{R}^+ \rightarrow H = \{A \subset \mathbb{R}^3 : A \text{ compact}\}$$

Examples may be given here by families of spline patches controlled by continuously moving control points $c_i(t)$, cf. Section 3.3.

6.1 A metric structure for medial modeling

In the previous setting, we compared radius functions only in the simplified special case in which they were defined on a common medial axis set. It is desirable to formulate the continuous change of the medial axis set together with the change of the radius function in a common topology. For this purpose, it is also possible to consider the preceding continuous deformation concept as a whole being describable within a general setting employing Hausdorff-topology and spaces of functions endowed with appropriate topologies. For this we define a metric on the product space built by the product of the two spaces $\tilde{H} \times F$, one of them being the above Hausdorff space

$$\tilde{H} = \{A : A \text{ compact subset of } \mathbb{R}^3 \cap B_k(0)\}$$

(\tilde{H}, d_H) being endowed with the Hausdorff metric d_H defined above. The other space F in the product $\tilde{H} \times F$ is given by the space of all continuous real-valued functions defined on the compact set $\tilde{B}_k(0)$. On the latter space of continuous functions we can define a metric

$$d_f(f, g) := \max_{x \in \tilde{B}_k(0)} \{|f(x) - g(x)|\}$$

for any pair of continuous real-valued functions f, g defined on $\tilde{B}_k(0)$.

In this context, it is quite important to understand that any continuous function defined on a compact subset $A \subset \tilde{B}_k(0)$ can be viewed as a restriction of an appropriately chosen function being continuous on all $\tilde{B}_k(0)$. This holds here since the space $\tilde{B}_k(0) \subset \mathbb{R}^3$ fulfills appropriate separation properties; see also T_4 axiom of Hocking and Young (1988) [2]. Clearly the metric on the product space is now defined by

$$d_{\tilde{H}}((A, r_1), (B, r_2)) := d_H(A, B) + d_f(r_1, r_2)$$

It can be shown that if

$$d_{\tilde{H}}((A_n, r_n), (A_0, r_0)) \rightarrow 0 \text{ then } d_H\left(\bigcup_{p \in A_n} B_{r_n(p)}(p), \bigcup_{p \in A_0} B_{r_0(p)}(p)\right) \rightarrow 0$$

The sequence of objects (each of which modeled by the union of discs) converges in the Hausdorff metric to the related limit object. Unions of discs obtained from members of a sequence of medial axis transforms converge against the discs union of the limit (medial axis transform). Clearly, if $\psi(t) = (A(t), r_t)$ is a continuous deformation path in $(\tilde{H} \times F)$ with (A_0, r_0) being the medial axis transform of a solid, then for the respective discs unions related to $\psi(t)$ we have Hausdorff convergence toward the solid corresponding to (A_0, r_0) .

1. However note that not every pair (A, r) will define a solid via the union $(\bigcup_{p \in A} B_{r(p)}(p))$
2. In case $(\bigcup_{p \in A} B_{r(p)}(p))$ defines a solid it may not have A as medial axis and $r : A \rightarrow \mathbb{R}$ may not be a maximal disc radius function.

Examples in the context of statement 1 above may be constructed easily in case we use a radius function r that may attain the value zero as then parts of the object might agree with the axis A that may be chosen deliberately wild. In case we assume that the radius function $r > 0$, then we may still have delicate situations in which the boundary of a domain obtained from a union of closed discs may at some points be locally homeomorphic to two arcs having tangential contact of a single point (see Figure 27). Figure 28 shows an example illustrating the claim in 2. Here $\bigcup_{p \in A} B_{r(p)}(p)$ defines a solid whose medial axis contains a topological circle while A does not.

6.2 Boundary representation in medial modeling

So far the Medial Modeling concept has been built on the idea to mold the solid by a union of discs. It may be preferable to represent the respective solid rather by appropriate

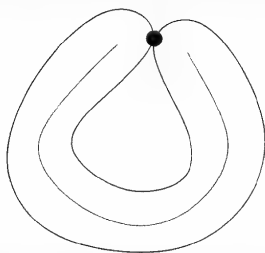


Figure 27. Tangential contact of envelopes. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

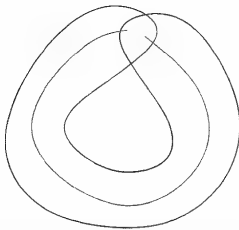


Figure 28. Self-intersection of envelopes. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

boundary surfaces cf. Section 4. The latter ones arise quite naturally in the medial modeling context. Here the boundary surface (curve) is created as the envelope surface of the family of discs belonging to the specific medial axis transform (see Figure 29). Let us assume that the medial axis is presented locally by a differentiable curve or surface patch being presented by parametric functions $m(u)$, with the radius function $r(u)$ depending on the parameter u as well.

It is possible to express the envelope surface using functions $env(u)$ in terms of expressions involving $m(u)$, $m'(u)$, $r(u)$, $r'(u)$. It is also possible to compute $env(u)$ and the curvature of the envelope curve. The latter computations need higher-order derivative information of the functions representing the medial axis and of the radius function; refer to Wolter and Friese (2000).

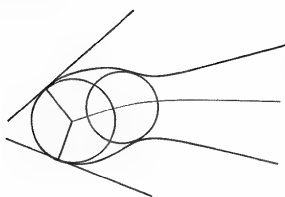


Figure 29. Construction of the envelope. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

Employing the concepts outlined above, different systems have been developed at the Welfenlab that can compute the envelope surface yielding a boundary representation of a solid whose medial surface and whose radius function have been given. More precisely the aforementioned medial modeling system computes for a parametric spline surface patch $m(u) : [0, 1] \times [0, 1] \rightarrow \mathbb{R}^3$ and for an associated radius function $r(u)$ the boundary surface of the corresponding solid whose medial surface is given by $m([0, 1] \times [0, 1])$ being a deformed rectangle embedded without self intersections into \mathbb{R}^3 , cf. Figure 30. Figure 31 illustrates the simplified special case in which $m : [0, 1] \rightarrow \mathbb{R}^2$ is a planar arc and the now 2-D solid corresponds here to a planar domain. At those positions where the center points of the maximal discs are located on the boundary of the medial patch, we get the related boundary surface of the solid from appropriate parts of maximal spheres. Here the construction (using the modeler) is valid if the normal

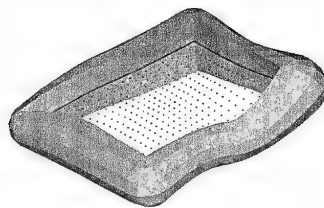


Figure 30. A medial patch inside its associated solid. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

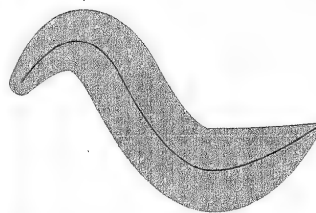


Figure 31. The 2-D case. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

segment (joining the point $env(u)$ of the envelope surface with the medial axis point $m(u)$) does not meet a curvature center of the point $env(u)$ of the envelope surface prior to meeting $m(u)$. Using the curvature formula for the envelope surface mentioned above and some additional criteria, it is easily possible to check if the radius function is admissible. This means that the previously stated curvature center condition must hold. Under those assumptions, it can be shown that the related envelope surface that we assume to be free of self intersections yields the boundary surface of a solid

$$\bigcup_{m(u) \in m([0, 1] \times [0, 1])} B_{r(u)}(m(u))$$

with $m([0, 1] \times [0, 1])$ being the medial axis of the solid being homeomorphic to a 3-D cube.

This result can be generalized to situations where the medial axis is built by a connected finite collection of patches. Again that collection of patches denoted by A will constitute the medial axis of a solid whose boundary surface is given by the envelope surface obtained via the disc radius function being defined on the collection of patches. Again we must assume that the envelope surface has no self intersections and that the above-mentioned curvature center assumption holds for the envelope surface.

The situation in which the medial axis is built by a collection of patches is far more complicated than the case in which the medial axis is given by a single patch. Therefore, we shall not go into a detailed discussion on this case in this survey paper. Suffice to say, in order to deal with that complicated case, envelope surfaces related to adjacent medial patches are joined along curves. The geometry of the intersection of medial surface patches, here the angles between intersecting medial patches at an intersection point, poses conditions that can be used to appropriately blend adjacent

envelope surfaces that are related to adjacent medial surface patches. These blended envelope surfaces are used to construct the boundary surface of a solid containing the aforementioned medial surface patches.

6.3 Medial modeling and topological shape

One of the major reasons why the medial axis is so important for the shape of a solid is because it essentially contains the homotopy type of a solid because it is a deformation retract of the solid; refer to Wolter (1992) and Wolter and Friese (2000). The following *topological shape theorem of the medial axis* applies: The Medial Axis $M(D)$ contains the essence of the topological type of a solid D .

Let ∂D be C^2 -smooth (or let ∂D be 1-D and piecewise C^2 -smooth, with $D \subset \mathbb{R}^3$). Then the Medial Axis $M(D)$ is a deformation retract of D , thus $M(D)$ has the homotopy type of D .

The proof of this theorem shows that it is possible to define a homotopy $H(x, t)$, as explained below the next figure, describing a continuous deformation process of the solid D . This deformation process depends on the time parameter t . The deformation starts with the solid. In Figure 32 this is a rectangle with a circular hole. During the deformation, points are moved along the shortest segments starting at the solid's boundary ∂D until the segments meet the dotted Medial Axis. The shortest segments are indicated by arrows in Figure 32.

We describe a homotopy

$$H(x, t) : (D \setminus \partial D) \times [0, 1] \rightarrow (D \setminus \partial D)$$

such that

$$H(x, 0) = x \quad \forall x \in D \setminus \partial D$$

$$H(x, t) = x \quad \forall x \in M(D)$$

$$H(x, 1) = R(x) \quad \text{with } R : D \setminus \partial D \rightarrow M(D) \setminus \partial D$$

For this we define the homotopy as follows:

$$H(x, t) := x + td(x, \psi(x)) \nabla d(\partial D, x)$$

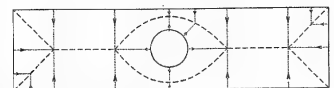


Figure 32. Deformation retract.

Here $d(x, y)$ denotes the function describing the distance between variable points x, y ; $\nabla d(x, y)$ describes the gradient of the distance function $d(x, y)$. $\psi(x)$ is defined as point where the extension of a minimal join from ∂D to $x \in (D \setminus \partial D)$ meets $M(D)$.

6.4 Medial modeling and meshing of solids

In the preceding section on medial modeling, we outlined geometrical concepts that were used to explain the deformation retract property stated in the topological shape theorem. We outlined also how to look at the solids boundary surface as an envelope surface that can locally be presented by a nonsingular parameterization map defined on the medial axis, (cf. Wolter and Friese (2000) for more details). All those geometric considerations immediately lead to insights explaining that the medial axis concept can be used nicely as to construct for the given solid a meshing partition that is naturally associated with the solid's medial axis.

We shall outline possibilities to use the medial axis for the meshing of solids by sketching some examples presented subsequently in several figures further down. In this context, some observations are relevant. In case the solid S is created with the medial modeler say with a medial axis being diffeomorphic to a square Q then we immediately obtain a quite simple parameterization of the solid. That parameterization map can be described by a differentiable map being defined on a solid PS containing all points in 3-space whose Euclidean distance to the unit square Q in the xy -plane is not larger than one. Here we identify the latter unit square with the parameter space of the medial axis surface. Our definition of the parameterization map is essentially obtained from the differentiable function $env(u)$ describing the envelope surface being the solid's boundary surface, (cf. Section 6.2). The respective parameterization map of the solid S maps an Euclidean segment in PS (joining any point u in the interior of Q orthogonally with the boundary of PS) linearly onto an Euclidean segment in S . The latter segment joins the medial axis point $m(u)$ with one of the two corresponding boundary points $env(u)$ in S . This segment in S meets the boundary surface of S orthogonally, (cf. Section 6.2). The outlined parameterization map of the solid yields a differentiable map f from PS onto S with a differentiable inverse f^{-1} . Figures 30 and 33 show the correspondence between the PS and S . In the simplified (lower dimensional) case, the solid S is a 2-D domain with its medial axis being now an arc instead of a 2-D surface. The maps f and f^{-1} can be used to map certain convex sets in PS onto convex sets in S . This can be used to get a partition of an approximation of the solid S into convex subsets. Figure 34 shows

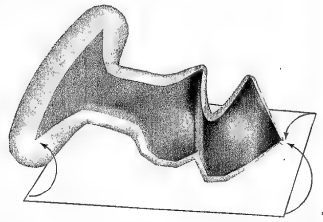


Figure 33. Medial axis of a solid. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

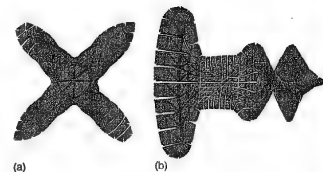


Figure 34. Meshes obtained from a medial axis representation. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

examples where fairly complicated engineering objects that have been modeled with our medial modeling system have now obtained tetrahedral meshes that have been constructed employing the geometrical concepts that were explained above.

7 ATTRIBUTES

Sometimes there is a need to store additional information associated with a geometric model. We have already seen such an example: topological information in a boundary representation, in this case adjacency information. There is a variety of other data that can be associated to a model, we will refer to all of this as *attributes of the model*. These can be attributes of physical origin, which alter the reception of the object by the user, or logical attributes, which relate the object to other objects or data. Physical attributes include photometric, haptical, and other material constraints, such as elasticity or roughness.

7.1 Textures

If attributes are quantifiable (which is true for most of the physical attributes), then they are often specified by textures, which are functions that relate surface points to certain quantities of the attribute. Formally a texture is defined by:

$$t: M \rightarrow V$$

where M is the set of points of the model and V is the set of possible values of the texture. M can consist either of the entire volume of the model or only the surface points depending on the nature of the attribute.

Normally textures are implemented using two maps

$$p: \mathbb{R}^2 \rightarrow M_S$$

and

$$v: \mathbb{R}^2 \rightarrow V$$

with p being the well-known parameterization of the surface points M_S . Then the texture t is given by

$$t := v \circ p^{-1}$$

Note that in practice one does not need to compute the inverse of p since the coordinates in parameter space (here identical to the texture coordinates) are known during the process of painting. Nevertheless, often it is rather difficult to find appropriate (i.e. nonsingular) mappings from the plane onto a given surface (in fact it is impossible as stated by the Hopf index theorem). This results in distortions of the texture near the singularities. To avoid this, one can use *solid textures* (cf. Peachey, 1985; Perlin, 1985). Here the map v is defined as

$$v: \mathbb{R}^3 \rightarrow V$$

and thus $t := v$ since $M \subset \mathbb{R}^3$. Note that while ordinary textures are implemented as pixel images, solid textures are represented by voxel spaces (cf. Section 2). This approach is slightly faster and avoids distortions induced by the parameterization, on the other hand it consumes far more memory. Furthermore, ordinary 2-D textures can often be easily derived from photographs whereas this is much more complicated for solid textures; see for example De Bonet (1997).

Solid textures are easily applied in areas, where the texture data itself result from real-world data, for example, a spatial scan of a material density or the like. Nevertheless, modeling a spatial texture can be intricate. The approach presented in Biwas, Shapiro and Tsukanov (2002) shows

how to combine traditional modeling techniques like CSG with the theory of distance functions to model arbitrary 3-D textures. For each textural attribute (here referred to as *feature*), its extremal sets are modeled as separate solids, then the gaps in between are filled via distance interpolation methods. A slightly different and more general approach can be found in Jackson *et al.* (2002) and Liu *et al.* (2003). These techniques are commonly referred to as *heterogeneous* or *inhomogeneous* modeling. The texture can then be kept in its quasicontinuous representation to benefit from the representations superior analytic properties, or it can be easily converted to a voxel space representation for faster rendering etc.

The most common use for textures are photometric textures, which are maps that modify the color of a surface point. Figure 35 shows a sphere, a photometric texture resembling marble and the sphere "wrapped" with the texture.

Note that the use of textures is not limited to color (although this is assumed widely in the literature), other common uses include *bump maps*, which are vector fields that alter the direction of the point normal (and thus altering the appearance of the surface locally near the point). Figure 36 shows a texture and the result of bump-mapping it onto a sphere. Note that not the sphere's geometry itself is changed but only the face normals.



Figure 35. Photometric marble texture. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

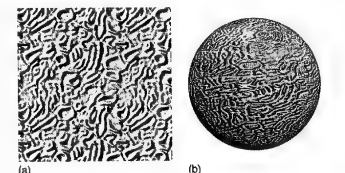


Figure 36. Bump-map on a sphere.

Textures are especially useful to model micro-scale aspects of surfaces, where detailed polyhedral modeling is too costly. For example, a micro-scale roughness of a stone surface can more efficiently be simulated by photometric and haptical textures than by subdividing the (macro-scale) smooth surface into tiny triangles. Furthermore, material properties like elasticity or particle density can be represented by 3-D textures. Rather than simulating the position of single, individually invisible particles, a quasicontinuous texture is applied to the model space.

7.2 Model parameters

A special class of attributes are the *model parameters*. As we have seen in the preceding chapters, most model representations have a set of parameters associated with them, for example, the set of control points for a spline patch. Sometimes it makes sense to view some of these parameters as attributes. Additional parameters can be added to most models, these include Euclidean motion matrices, stiffness constraints, and the like. It is then sometimes more appropriate to allow these parameters to change over time, making them effectively attached *functions* rather than constants. These techniques lead to the theory of *physics-based modeling*, refer to Metaxas (1997) for a comprehensive overview of this topic.

7.3 Scripts

An example of logical attributes are scripts. These are parts of code or methods that can be invoked when certain constraints of the model are met. For example, a 3-D object can have scripts attached that react to user interaction, when the object is selected in an interactive scene. Another example would be a script that is activated on collisions of the object with other parts of the scene. Scripts are especially useful in applications like physical modeling, where the modification of one object may require also modifications to associated objects. Rather than attributing these dependent objects to the calling object passively and letting the main program do the work, the tasks are carried out directly by the objects involved.

The idea of scriptable attributes has now been around for several years without finding a broad acceptance. Nevertheless there have been some prototype implementations like the *Odyssey Framework* (cf. Brockman *et al.*, 1992; Cobourn, 1992).

8 OUTLOOK AND CONCLUDING REMARKS

A theoretically and practically difficult topic that we barely touched upon in this paper considers aspects related to the analysis and computation of singularities of geometric loci. Those singularities may come up on various occasions. They quite often concern the structure of geometrically defined solutions of nonlinear equations being crucial to define precisely the local and global topological structure of solids and their parts. Those singular sets very naturally come up, for example, when we are dealing with surface intersections that may be related to Boolean operations carried out for solids bounded by surfaces (cf. to Kriezis, Patrikalakis and Wolter, 1992; Patrikalakis and Maekawa, 2002). Similar problems also cause major difficulties in the context of CSG modeling; see Section 5, (cf. Hoffmann, 1989). Simply spoken, whenever a set under consideration has not the structure of a topological or of a differentiable manifold, then it will have a singular structure at some locations. For an important class of singular sets, this can be rephrased by saying singular sets in the Euclidean space cannot be represented by solutions of equations corresponding to some differentiable functions whose differential has a maximal rank at all points belonging to the 'singular set' under consideration. Computations and constructions related to the medial axis in Section 6 of this paper often contain as their most difficult part computations and representations of the singular subsets of the medial axis (cf. Section 5). In general, analyzing and understanding the mathematical structure of singular sets is sometimes quite difficult and may require the use of sophisticated and fairly advanced mathematical methods related to singularity theory. The mathematical and computational trouble caused by mathematically singular sets is enhanced by an additional fundamental problem in this context. One of the crucial difficulties that we encounter in geometric modeling is caused by the fact that all our models are usually represented in a discrete space and they only use points on a finite 3-D grid having a limited resolution. This implies that even in cases where we are dealing with solids, bounded by surfaces consisting of triangular facets only, we still may have difficulties carrying out Boolean operations. Those difficulties are caused by the fact that for certain geometric configurations we cannot properly compute the intersection set of two triangular facets. The latter problem may result in a (wrong) decision assuming the intersection point to be wrongly inside or outside of some triangular facets. In the end, all this may contribute to major topological inconsistencies and contradictions causing a failure of the system. In our view, the state of the art in geometric modeling

related to all of the aforementioned areas still needs substantial improvements by innovative concepts. Those new concepts to be developed should benefit from ideas inspired by advanced mathematical concepts from computational differential geometry and from singularity theory; cf. the pioneering work of the late Thom (1975), Bruce and Giblin (1992), Arnold, Gusein-Zade and Varchenko (1985), and Arnold (1990). New exciting research by Leymarie (2003) uses the medial axis concept (cf. Section 6) in combination with ideas resting on a singularity analysis of distance wave fronts as to develop new methods for 3-D shape representation that are applicable in a context of discrete point sets.

Another currently very active area related to geometric modeling is dealing with data compression. Often huge amounts of data points may arise from measurements or from construction procedures, for example, when large objects are constructed by many patches. Those collections of many patches need to be simplified and reduced, that is, approximated by a surface whose description needs far less data (cf. Bremer *et al.*, 2001). However this approximation often must fulfill some specified accuracy requirements, for example, concerning placement. Furthermore, quite often we must meet some topological conditions such as that the approximating surface may not have self intersections and singularities. Data corresponding to evaluation of continuous functions defined on geometric 2-D or 3-D objects may be obtained by measurements or by time consuming computational procedures such as those used in the area of differential equations. In all these cases, one may encounter extremely large data sets that are far beyond the size that can be handled on current computers. In those situations, one appreciates good approximation methods allowing an efficient approximation of the given data (or of that respective function). The description and evaluation of the approximation should need far less data than the original data set. Furthermore it should be possible to process the approximation data (substituting the original ones) efficiently on the computer for the particular computational purpose. This survey paper here has been touching the basics of related topics, for example, in the Sections 3.3 and 3.5. Suffice to say that new concepts of wavelet and multiresolution theory appear to provide powerful tools that currently drive the progress in the respective fields that may be considered to belong to the subject of data compression (cf. to Mallat, 1998; Stollnitz, DeRose and Salesin, 1996). It should be mentioned that very recent innovative efforts in the area of data compression employing new methods from a so-called discrete Morse theory benefit from concepts that have been developed in the classical areas of modern global differential geometry and differential topology (cf. Edelsbrunner, Harer and Zomorodian, 2001; Milnor, 1967). Meanwhile, there even exists a new field called

'computational topology' presenting fundamental research for geometric modeling that has been inspired strongly by methods and questions and ideas stemming from the classical area of topology and differential topology, for example, refer to the recent work by Amenta, Peters and Russell (2003).

Historically, geometric modeling has been developed as a basic science for Computer Aided Design. In its early days, the latter field has been employing descriptive geometry and Bezier geometry to design the shape of objects electronically instead of using blueprints created in technical drawings with the help of compasses and ruler. Meanwhile, engineers want computer aided modeling systems whose capabilities go far beyond Computer Aided Design. Those systems shall not only describe the shape of objects but should allow also the simulation of various physical properties of the design object. This essentially implies that the computer system must be capable to solve partial differential equations (PDEs) being defined on the geometry of the designed object. For this purpose, we may need systems allowing very rapidly (ideally in real time) a good automated meshing procedure of the geometric design object. The resulting mesh must be appropriate for the approximate solution of the respective PDE used to analyze some properties of the designed object. Future geometric modeling and meshing systems will have to address those important needs. Those systems may therefore integrate the design and meshing functionalities in combined systems as it has been, for example, suggested in our medial modeler system described in Section 6.4. In order to handle the combined needs of designing shape as well as designing the physics of objects, it appears to make sense that the different engineering communities doing geometric modeling research, meshing research, and computational engineering (PDE) research will cooperate more closely in the future. This collaboration should initiate learning processes in which each community should profit from the knowledge available in the other communities.

Overall, assessing future developments, we think that new developments in geometric modeling and also in the aforementioned areas will increasingly employ concepts and insights from singularity theory, from local and global differential geometry, and from advanced (singular) wavelet theory. The latter areas will help to provide mathematical concepts and tools being relevant to analyze and compute delicate singularities that may, for example, be encountered analyzing dynamical processes related to various types of PDEs defined in the context of a physical analysis of the design object.

Finally we present a remark that corroborates our statement that new developments in geometric modeling and related fields benefit from using synergistically advanced

concepts from global and local differential geometry. Geometric modeling is primarily involved with shape construction but it is dealing also with the area of shape interrogation and by that geometric modeling is related to shape cognition of 3-D and 2-D objects. Shape cognition is concerned with methods identifying automatically the shape of an object in order to check if the shape design is already in a database containing shape design models being, for example, protected by some copyright. We want to point out that recent advances on new strong methods concerning shape cognition benefit also from advanced concepts of global and local differential geometry such as singularities of principal curvature lines called umbilics (cf. Maekawa, Wolter and Patrikalakis, 1996; Ko *et al.*, 2003a; Ko *et al.*, 2003b).

ACKNOWLEDGMENTS

The authors like to thank Philipp Blanke and Patrick Klie being team members of the Weitenlab for carefully proof reading this paper.

NOTES

- [1] Topologists call $f[D]$ the image set of f and \mathbb{R}^3 the range of the map. Analysts often call $f(D)$ the range of the map and do not introduce a special name for the right-hand side of f .
- [2] This consideration shows that any radius function being a continuous function on a compact subset A of $\bar{B}_R(0)$ is restriction of some continuous function defined on the set $\bar{B}_R(0) \supset A$.

REFERENCES

- Amenta N, Peters TJ and Russell AC. Computational topology: ambient isotopic approximation of 2-manifolds. *Theor. Comput. Sci.* 2003; 305(1–3):3–15.
- Arnold VI. *Singularities of Caustics and Wave Fronts*. Kluwer Academic Publishers: Dordrecht, 1990.
- Arnold VI, Gusein-Zade SM and Varchenko AN. *Singularities of Differentiable Maps*, vol. I, II. Birkhauser: Boston, 1985.
- Baumgart B. A polyhedron representation for computer vision. In *National Computer Conference*, AFIPS Conference Proceedings, Anaheim, 1975; 589–596.
- Biswas A, Shapiro V and Tsukanov I. Heterogeneous Material Modeling with Distance Fields. *Comput. Aided Geometr. Des.* 21(3);, 2004, 215–242.
- Blum H. Biological shape and visual science. *J. Theor. Biol.* 1973; 38:205–287.
- Bremer PT, Hamann B, Kreylos O and Wolter F-E. Simplification of closed triangulated surfaces using simulated annealing. In *Proceedings of the Fifth International Conference on Mathematical Methods for Curves and Surfaces*, Oslo, July 2000, *Mathematical Methods in CAGD*, Vanderbilt University Press: Tennessee, 2001; 45–54.
- Brockman JB, Cobourn TF, Jacome MF and Director SW. The odyssey CAD framework. *IEEE DATC. Newsletter on Design Automation*, 1992; 9(2):91–93.
- Bruce JW and Giblin PJ. *Curves and Singularities* (2nd edn). Cambridge University Press, 1992.
- Cobourn TF. Resource Management for CAD Frameworks. *Dissertation*, Carnegie Mellon University, May 1992, CMUCAD-92-39.
- De Bonet JS. Multiresolution sampling procedure for analysis and synthesis of texture images. In *ACM SIGGRAPH, ACM Conf. Proc.*, Los Angeles, 1997.
- Edelsbrunner H, Harer J and Zomorodian A. Hierarchical morse complexes for piecewise linear 2-manifolds. In *Symposium on Computational Geometry*, Medford, 2001.
- Farin G. *Curves and Surfaces for Computer Aided Geometric Design: a Practical Guide* (3rd edn). Academic Press: San Diego, 1993.
- Foley JD, van Dam A, Feiner SK and Hughes JF. *Computer Graphics: Principles and Practice* (2nd edn in C). Addison-Wesley: Reading, 1996.
- Hocking JG and Young GS. *Topology* Dover Publications Inc.: New York, 1988.
- Hoffmann CM. *Geometric and Solid Modeling: An Introduction*. Morgan Kaufmann: San Mateo, 1989.
- Hoschek J and Lasser D. *Fundamentals of Computer Aided Geometric Design*. A K Peters, Wellesley, 1993.
- Jackson TR, Cho W, Patrikalakis NM and Sachs EM. Analysis of solid model representations for heterogeneous objects. *ASME Trans. JCISE: J. Comput. Inf. Sci. Eng.* 2002; 2(1):1–10.
- Ko KH, Maekawa T, Patrikalakis NM, Masuda H and Wolter F-E. Shape intrinsic properties for free-form object matching. *ASME J. Comput. Inf. Sci. Eng. (JCISE)* 2003b; 3(4):325–333.
- Ko KH, Maekawa T, Patrikalakis NM, Masuda H and Wolter F-E. Shape intrinsic fingerprints for free-form object matching. In *Proceedings of the Eighth ACM Symposium on Solid Modeling and Applications*. Seattle, WA, June 2003b; 196–207.
- Koenderink JJ. *Solid Shape*. MIT Press: Cambridge, 1990.
- Kriezis GA, Patrikalakis NM and Wolter F-E. Topological and differential-equation methods for surface intersections. *Comput. Aided Des.* 1992; 24(1):41–55.
- Leymarie FF. *Three-Dimensional Shape Representation via Shock Flows*. PhD thesis, Brown University, Division of Engineering, Providence, 2003.
- Liu H, Maekawa T, Patrikalakis NM, Sachs EM and Cho W. Methods for feature-based design of heterogeneous solids. *Comput. Aided Des.* 2003, (unpublished, available on the www).
- Lorensen W and Cline H. Marching cubes: a high resolution 3D surface construction algorithm. *ACM/SIGGRAPH Comput. Graph.* 1987; 21(4):163–169.
- Mallat SG. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 1998.
- Maekawa T, Wolter F-E and Patrikalakis NM. Umbilics and lines of curvature for shape interrogation. *Comput. Aided Geometr. Des.* 1996; 13(2):133–161.
- Mäntylä M. *An Introduction to Solid Modeling*. Computer Science Press: Rockville, 1988.
- Metaxas DN. *Physics-Based Deformable Models: Applications to Computer Vision, Graphics and Medical Imaging*. Kluwer Academic Publishers: Boston, 1997.
- Milnor J. *Morse Theory*. Princeton University Press: Princeton, 1967.
- Nowacki H, Bloor MIG and Oleksiewicz B. *Computational Geometry for Ships*. World Scientific: Singapore, 1995.
- Patrikalakis NM and Maekawa T. *Shape Interrogation for Computer Aided Design and Manufacturing*. Springer: Berlin, 2002.
- Peachey DR. Solid texturing of complex surfaces. In *SIGGRAPH 85*, San Francisco, 279–286.
- Perlin K. An image synthesizer. In *SIGGRAPH 85*, San Francisco, 287–296.
- Piegl I. and Tiller W. *The NURBS Book*. Springer: Berlin, 1995.
- Samet H. The quadtree and related hierarchical data structures. *ACM Comput. Surv.* 1984; 16(2):187–260.
- Stollnitz EJ, DeRose TD and Salesin DH. *Wavelets for Computer Graphics: Theory and Applications*. Morgan Kaufmann, San Francisco, 1996.
- Thom R. *Structural Stability and Morphogenesis: An Outline of a General Theory of Models*. Benjamin-Cummings Publishing: Reading, 1975.
- Wolter F-E. Distance function and cut loci on a complex riemannian manifold. *Arch. Math.* 1979; 32(1):92–96.
- Wolter F-E. *Cut Loci in Bordered and Unbordered Riemannian Manifolds*. PhD Dissertation, Technical University of Berlin, Department of Mathematics Berlin, 1985.
- Wolter F-E. *Cut Locus and Medial Axis in Global Shape Interrogation and Representation*. MIT Seagrant Report, 1992, National Sea Grant Library, NSGL#: MIT-T-93-002, Program#: MITS93-11.
- Wolter F-E and Friese K-I. Local and global geometric methods for analysis interrogation, reconstruction, modification and design of shape. In *Proceedings of CGI 2000* Geneva, 2000; 137–151.
- Yamaguchi F. *Curves and Surfaces in Computer Aided Geometric Design*. Springer: Berlin, 1988.

Chapter 17

Mesh Generation and Mesh Adaptivity

P. L. George¹, H. Borouchaki², P. J. Frey¹, P. Laug¹ and E. Saltel¹

¹INRIA, Projet Gamma, Domaine de Voluceau, Rocquencourt, Le Chesnay Cedex, France

²Université de Technologie de Troyes, Troyes Cedex, France

| | |
|--|-----|
| 1 Introduction | 497 |
| 2 A Brief History | 498 |
| 3 Mesh-Generation Methods | 499 |
| 4 Quality Meshing and Adaptivity | 502 |
| 5 Adaptive FEM Computations | 510 |
| 6 Large-size Problem, Parallelism and Adaptivity | 516 |
| 7 Meshing for Moving Boundary Problems | 517 |
| 8 Application Examples | 519 |
| 9 Conclusions | 520 |
| References | 521 |
| Further Reading | 523 |

1 INTRODUCTION

To carry out a finite element analysis (FEA or FEM), or any type of analysis such as a boundary element method (BEM) or a finite volume method (FVM), which requires the use of a spatial decomposition of the domain of interest, it is necessary to construct an appropriate mesh of the corresponding computational domain. This is the first step we face when using such methods. There are various automatic mesh-generation methods that are widely used in software packages (Frey and George, 2000). Nevertheless, these methods are subject to rapid changes and improvements, and the demand in terms of meshing

facilities is also constantly changing. At present, developing quality meshes adapted to the physics of the problem to be solved represents a field of intensive research (Thompson *et al.*, 1999).

Meshes can be categorized as being *structured* or *unstructured*. Structured meshes follow a simple structure like a *grid* or a *finite difference network* at the vertex connectivity level. Construction methods for such meshes are purely *algebraic* (Cook, 1974), of the *multiblock* type (Allwright, 1988) or based on appropriate partial differential equations (*PDE solutions*) (Thompson *et al.*, 1985). Geometries that can be successfully handled are more or less close to quadrilateral (hexahedral) regions or decomposed by means of such simply shaped regions. Arbitrarily shaped domains are more tedious to consider in this way. Therefore, unstructured meshes appear to be the solution. Construction methods, in this case, fall into three categories: (based on) *hierarchical spatial decompositions* (quadtree, octree) (Yerry and Shephard, 1984), (on) *advancing-front* strategies (Van Phai, 1982; Lo, 1985; Löhner, 1997) and (by means of) *Delaunay-type* methods (Weatherill and Hassan, 1994; Marcum and Weatherill, 1995; George and Borouchaki, 1998).

Hierarchical decomposition (or tree-based) methods start with a unique parent (or root), a cell enclosing the entire domain, and they split this single cell into four (eight) similar (or congruent) subcells according to a given criterion. A tree is thus formed and analyzed before being recursively subdivided (according to the above criterion), until certain properties hold. The resulting tree then allows the mesh elements to be defined, and the exterior elements are removed. *Advancing-front*-based methods use the discretization of the domain boundaries as an initial front. Each edge (triangular facet) in this front is then used to construct a triangle

(a tetrahedron). A new front is then defined and dealt with in the same way. The mesh is completed when the current front is empty. It is worth remarking that this approach allows for quadrilateral mesh construction (in two dimensions), but is rather tedious to extend to three dimensions (for hexahedral mesh construction). *Delaunay*-type methods make use of Delaunay triangulation algorithms. Such an algorithm results in the insertion of a point in a given mesh and completes a mesh of the convex hull of the set of points formed by the vertices of the domain boundary discretization (the input data). This material can be used as a point-to-point connector for mesh construction. Once the boundary points have been inserted, the boundary discretization can be obtained (a rather tedious task, at least in three dimensions, whereas in two dimensions, a series of edge flips allows a solution), and points are created inside the domain before being connected in turn. The final mesh is obtained when the domain is saturated (in the sense that it is no longer necessary to add new points).

The above discussion concerns the so-called *classical* mesh-generation problem. Given a discretized boundary, we complete at best, a mesh in the corresponding domain composed of elements that are judged to be reasonable in terms of *size* and *quality* (shape). The size is related to the greater or lesser thickness of the boundary discretization and remains to be properly defined inside the domain. The quality is related to the element aspect ratios. Moreover, the way in which these two parameters vary in a given region is an important issue. It could be observed that these criteria greatly depend on the way in which the field points are created, such an issue being somewhat tedious. Turning to mesh *adaptation* involves looking at the same problem while adding a series of constraints about element size and element directions (*anisotropic* case), (Peraire *et al.*, 1987, 1992; D'Azevedo and Simpson, 1991; Vallet, 1992). The aim is no longer to obtain the best possible mesh but to complete elements of a prescribed size and direction where necessary. *Tree-based* methods are unlikely to be suitable in this respect (in particular for handling nonuniform directional constraints). *Advancing-front* or *Delaunay*-type methods promise greater flexibility and so are more likely to be suitable.

This chapter is made up of 9 sections. Section 1 provides a general introduction to the various problems in hand. Section 2 gives a brief history about mesh-generation methods. Section 3 gives an overview of the most popular mesh-generation methods, with a special emphasis on automated methods. Section 4 is the core of the chapter and contains a number of issues related to adaptation methods. We focus on a Delaunay-type method and propose a general formulation for a mesh-generation problem. More precisely, we introduce the notion of a *regular* mesh. We

then demonstrate how this framework allows a formulation of the mesh-generation problem for adapted meshes. Then, we give a construction method based on these ideas. The parametric together with the discrete curve and surface cases are discussed in separate subsections. Finally, the volume-meshing aspect is discussed. Section 5 establishes the relationship between adapted mesh-generation methods and adaptive FEM (finite element method) simulations, after which, large-size mesh creation is discussed, leading to parallel meshing (Section 6). Moving boundary meshing problems are addressed in Section 7. To demonstrate how the previous approaches work, Section 8 shows a series of application examples in different engineering disciplines, together with other more or less derived applications. Finally, Section 9 gives some conclusions and mentions what the future may hold.

2 A BRIEF HISTORY

Without fear of contradiction, it can be claimed that the finite element method (FEM) was pioneered by engineers and practitioners in the early fifties (see earlier editions of Zienkiewicz, 1977 for instance). Then, during the sixties, mathematicians established the theoretical and mathematical foundations of this method (see Ciarlet, 1991 or Hughes, 1998 among many other references). The FEM then became widely used by various categories of people. A number of applications in different fields of engineering motivated the development of mesh-generation methods. Except when considering a unit square region or simply shaped geometries where the mesh generation is straight forward, concrete applications in arbitrary domains required designing and implementing automated mesh-generation methods. A pioneering work by George (1971), in the early seventies, demonstrated an advancing-front method for two-dimensional geometries. Quadtree-based mesh-generation methods were initiated by Shephard (Yerry and Shephard, 1983) a few years later. Delaunay-based mesh-generation methods were introduced by various authors (Hermeline, 1980; Watson, 1981).

As for three dimensions, computer facilities available from the eighties (including memory capacity and CPU efficiency), together with the need for more realistic simulations, triggered the investigation of mesh-generation methods capable of constructing three-dimensional meshes. In this respect, advancing-front, octree-based and Delaunay-type methods were extended to this case, while surface meshing received particular attention. First references include Hermeline (1980), Watson (1981), Yerry and Shephard (1984), Löhner and Parikh (1988), Joe (1991), Weatherill and Hassan (1994), and Marcum and Weatherill

(1995). Nowadays, intensive research and advanced developments are the topics of a number of groups throughout the engineering community.

Current authoritative literature about mesh-generation methods includes a number of monographs (Thompson *et al.*, 1985; George, 1991; Knupp and Steinberg, 1993; Carey, 1997; George and Borouchaki, 1998; Frey and George, 2000), a handbook (Thompson *et al.*, 1999), together with papers in 'specialized' annual conferences, and various survey papers.

3 MESH-GENERATION METHODS

Specific geometries can be handled by means of specific methods. Domains of a peculiar shape can be dealt with, taking advantage of their specificities. Convex domains (with no hole) more or less close to a (deformed) quadrilateral (hexahedron) can be dealt with using *algebraic* methods (Cook, 1974) or a PDE-based method (Thompson *et al.*, 1985). Arbitrary domains can be decomposed using a number of such simply shaped regions and then, *multiblock* methods allow for a solution (Allwright, 1988). Nevertheless, splitting a domain in this way is a rather tedious task and thus one that calls for fully automated methods that fall into three categories. The first type makes use of a hierarchical spatial decomposition that results in constructing a tree, which, in turn, allows the mesh elements to be created. The second type of method can be seen as a greedy algorithm, where a piece of the domain that has not yet been meshed is covered by a mesh element, thus resulting in a 'smaller' void region to be considered. Methods of the third type involve Delaunay triangulation algorithms revisited so as to produce meshes (and not only triangulation). The following gives a brief description of these three types of methods.

Planar and volume domains

Before entering into this description, we give an indication of what we term as a *mesh-generation problem*. We are given a domain Ω in \mathbb{R}^2 or \mathbb{R}^3 . This domain is defined by its boundary, Γ , which, in turn, is known by means of a discretization. The latter is a collection of line segments in two dimensions, these segments defining a polygonal approximation of Γ . In three dimensions, the boundary is defined by means of a triangulated surface (a list of triangles or quadrilaterals) which, again, defines an approximation of Γ . The problem is then, using this sole data, to construct an appropriate mesh of Ω (actually, an approximation of Ω). Such a context defines a so-called *classical* mesh-generation problem in which the goal is to recover Ω by quality elements where the notion of quality only refers

to aspect ratio (or element shape), mesh gradation but does not explicitly include any information about element sizing. Specifying sizing or directional requirements leads to a so-called *adapted* or *controlled* mesh-generation problem where the targeted elements, as above, must be well shaped, nicely graded and, in addition, must conform to a given sizing or directional specification (referred to as a metric in the following). The description below mainly concerns a classical mesh-generation problem, while a large part of the remaining sections discuss the other mesh-generation issue (i.e. how to complete an adapted mesh and therefore access adaptive computations).

Surface domains

Surface domains are different from planar or volume domains in the sense that a surface mesh must conform to the geometry of the surface in hand. In other words, smoothness, quality, and gradation are demanded, but it is mandatory to match the geometry. This leads to paying particular attention to the surface curvatures, thus making a purely two-dimensional approach unlikely to be suitable. In short, there are two ways to define a surface. A parametric definition involves defining a parametric space (thus in two dimensions), together with an appropriate mapping function. A discrete definition allows the geometry to be described by means of a mesh which is, *a priori*, a *geometric* mesh, and not a *computational* mesh. To some extent, parametric surfaces can be meshed using a two-dimensional method, provided information about the geometry is used to govern the method. On the other hand, discrete surfaces must be considered using the geometric properties of the surface directly.

3.1 Quadtree–octree based methods

Quadtree (octree)-based mesh-generation methods are methods whereby the domain is covered by elements using a recursive subdivision scheme based on a spatial tree structure. A given discretization (a surface mesh) of the boundary of the domain is first enclosed in a square (a cube). This initial box (the root of the tree structure) is then recursively subdivided into four (eight) similar sub-boxes, until a certain consistency is attained between the boxes and the boundary items. This recursive scheme can also be envisaged as a tree-decomposition procedure. Once the decomposition has been achieved, mesh elements are generated by subdividing tree cells in a conforming way (using predefined patterns). An optimization stage is usually required as a final step as the intersections between tree cells and the boundary items may lead to the creation of ugly-shaped elements.

Tree-based methods are very versatile and are able to handle complex geometries. However, unlike advancing-front or Delaunay-based methods, the original boundary discretization is usually not preserved in the final mesh of the domain. A variant of this approach consists in considering that the domain boundary is known via a geometric modeling system (no input boundary discretization is supplied). The insertion of boundary items in the tree structure is then performed through geometric queries.

Tree construction

The general scheme for constructing a spatial decomposition of a computational domain was originally proposed by Yerry and Shephard (1983). It consists in two successive steps: (i) the construction of a spatial covering up from a bounding box of the domain and (ii) the creation of internal vertices and mesh elements. It can be seen as an incremental method that results in inserting a boundary item into a current spatial decomposition. Assuming that the domain boundaries are known via a discretization, an initial cell is created corresponding to the bounding box of the domain. All boundary items are inserted into the cells of the current decomposition with respect to their dimension (from points to faces).

Let A be the current tree structure after the insertion of n boundary entities and let E be the next entity to be inserted. The cell C containing E is identified (here the tree structure proves useful for searching purposes). If C is empty (no other entity has been associated with it), then E is associated with C , if not, cell C is subdivided into equally sized subcells (four in two dimensions and eight in three dimensions) and the process is repeated into these subcells. At completion, a tree structure is defined, in which the depth (the maximum level of refinement) corresponds to the minimal distance between boundary items. In order to simplify the creation of elements, this tree structure is balanced so as to reduce the size ratio between adjacent cells to a factor of two at most.

Mesh-element construction

The creation of internal mesh vertices and mesh elements is straightforward. All internal cell corners are considered as mesh vertices. The tree-balancing rule leads to a substantial reduction in the number of possible configurations for a given cell, as compared with its neighbors. Therefore, a predefined pattern is associated with each internal cell configuration in order to provide a conforming mesh of the cell (16 such templates exist in two dimensions and 78 in three dimensions). The boundary cells (intersected by the boundary discretization) can be treated in a similar way. The mesh of the bounding box of the domain is then obtained by merging all the elements created at the cell

level. To obtain the final mesh of the domain, a coloring procedure is used to remove external elements. Obviously a mesh-optimization procedure must be applied to get rid of the badly shaped elements that may have been created because of cell/surface intersections, and to improve the overall mesh quality.

Surface meshing versus octree methods

A variant of the octree decomposition consists in building a surface mesh from an analytical surface representation, usually known via a geometric modeler. The depth of the tree can indeed be related to edge length h and to the size b of the bounding box by the relation $p = \log_2(b/h)$. As edge size h is proportional to the minimum of the principal radii of curvature, denoted by ρ , a lower bound for the tree depth can be found:

$$p \geq \log_2 \left(\frac{b}{\alpha \rho} \right)$$

where α is related to the geometric approximation. The local intrinsic properties of the surface (principal radii of curvature, vertex normals, etc.) can be used to construct the tree. In practice, the cells are refined until the geometric approximation criterion is satisfied.

3.2 Advancing-front methods

Advancing-front methods take the given boundary discretization (mesh) and create the mesh elements within the domain, advancing in from the boundary until the entire domain has been covered with elements. Such an element is constructed on the basis of one entity (edge or triangle) of a so-called front (initiated by the boundary items), which is connected with a node appropriately chosen among the existing nodes or created according to some quality criteria. Once an element has been formed, the front is updated and the process is repeated until the current front is empty.

Advancing-front methods were primarily developed to handle planar and volume domains but can be used to construct surface meshes.

Point creation, point connection

Formally speaking, the advancing-front procedure is an iterative procedure that attempts to fill the as-yet unmeshed region of the domain with elements (George, 1971). At each step, a front entity is selected, and a new (optimal) vertex is created and inserted in the current mesh, if it forms a new well-shaped element. The boundary discretization should be orientable. The mesh elements are created on the basis of entities (edges or faces) of the current front (the initial front being the boundary discretization). Central to

the advancing-front technique, the creation and insertion of an optimal point from a selected front entity requires some care. At first, an optimal point is computed (resulting in the creation of an optimal element). This point is then checked against neighboring vertices in the current mesh. A candidate point and the corresponding virtual element is analyzed to see whether it intersects the existing mesh elements or front entities. A candidate element is a valid element if none of its edges (faces in three dimensions) intersect any front entity and if it does not contain any mesh entity (for instance a vertex or an element). Once a candidate point has been retained, it is inserted into the mesh, the relevant mesh element is created, and the front is updated. At completion, a mesh-optimization procedure is applied to locally improve the element shape quality.

Notice that in three dimensions, some nasty configurations may occur, which prevent the creation of a valid mesh. In such cases, it can be useful to remove the last created elements and to restart the procedure while changing some parameters to avoid encountering the same problem again.

Surface meshing versus advancing-front methods

The same concept can be applied to create surface meshes. The main difference lies in the iterative algorithm used to find an optimal point location, given a front edge. Given an edge AB , the optimal point P is computed so as to construct an optimal triangle on the surface. The third point P is determined using an angle α between the stretching direction and the tangent at the midpoint M of AB (or using an average tangent plane and setting α to zero). This point does not necessarily belong to the surface as it has been created in a local tangent plane. If the surface is known via a geometric modeler, a query can provide the closest point from P onto the surface. Candidate points are identified as those lying in the disk of center at P and radius $\kappa \times \delta$, where κ is a positive coefficient and δ denotes the radius of a region nearby P . The size of the triangle is locally adapted to the surface curvature, that is, it is proportional to the minimal radius of curvature.

3.3 Delaunay-type methods

Many people find Delaunay-type methods very appealing, as the keyword Delaunay has a touch of elegance, based on various theoretical issues (see Preparata and Shamos, 1985 and Boissonnat and Yvinec, 1997) about properties of such triangulations. Despite this, a more subtle analysis and actual experience both indicate that those theoretical issues are unlikely to be usable in the context of creating FE meshes. Nevertheless, Delaunay triangulation algorithms can be revisited so as to be included in the ingredients of the so-called *Delaunay mesh-generation methods*.

Delaunay triangulation

While alternative methods exist, a popular and straightforward method for generating a Delaunay triangulation of a given set of points in \mathbb{R}^2 or \mathbb{R}^3 is the so-called *Bowyer-Watson algorithm*, also developed by Hermeline at the same time (1981). It is an incremental method that results in inserting one point in a given Delaunay triangulation.

Let T be the Delaunay triangulation of the set of n points and let P be a point to be inserted; under some realistic assumptions, this method simply reads $T = T - C(P) + B(P)$, where $C(P)$ is the cavity associated with point P and $B(P)$ is the remeshing of $C(P)$ based on P . Cavity $C(P)$ is the set of simplices in T whose open circumdisk (circumball in three dimensions) contains point P , while ball $B(P)$ is simply the simplices constructed by connecting P with the boundary edges (or triangles) of $C(P)$. Theoretical issues show that provided the former T is Delaunay, the resulting T with P as a vertex is also Delaunay. This method allows planar and volume triangulation to be generated (indeed it can readily be applied to any number of dimensions), while it is meaningless for surface triangulation (of a general variety).

Variations of this incremental algorithm include constrained versions, where the constraints are of a topological nature (e.g. specified edges (or triangle facets) are maintained through the process) or of a metric nature (e.g. additional properties such as element quality control are included in the construction). Also, non-Delaunay triangulations can be successfully carried out provided an adequate construction of $C(P)$ and anisotropic cases can be addressed.

Delaunay-based meshing algorithms

The above-mentioned constrained incremental method can be used as part of a meshing algorithm (thus referred to as Delaunay-based). Let Ω be the domain to be meshed and let Γ be a discretization of its boundary. The set of points in this boundary discretization is triangulated using the above incremental method (after being included in a convex bounding box). The resulting triangulation is a triangulation of the introduced box where, in general, extracting a mesh of Ω is unlikely to be possible. This is due to the fact that the boundary entities are not necessarily edges or facets of this triangulation. In other words, inserting the two endpoints of a boundary edge (the three vertices of a boundary triangle) may result in a triangulation where this edge (triangle) does not exist.

Various methods have been developed to allow for the regeneration of such missing entities. In two dimensions, edge-swapping operators result in what is needed, while in three dimensions, the same is rather tedious. Nevertheless,

those methods readily work and result in a triangulation of the box where all the boundary entities are established. As a consequence, a mesh of Ω can be obtained by suppressing those elements in the box outside the domain.

At this time, we have in hand a mesh of Ω whose vertices are, roughly speaking, the boundary vertices. Such a mesh is unlikely to be suitable for FE computations. Therefore, field points must be created before being inserted. Methods to create the necessary field points include using the centroid or the circumcenter of some mesh elements on the basis of appropriate criteria (such as density requirement, element quality concern, etc.). Other methods make use of a tree structure to define those field points or introduce points along the edge elements (for any dimensions) or, again, use an advancing-front strategy to locate those points. At completion, for example, when the mesh is saturated (according to the selected criteria), the resulting mesh is optimized (with regard to quality measures) by means of edge or facet swapping, node repositioning, and so on, as a Delaunay triangulation is not, in general, and particularly in three dimensions, a quality triangulation (for FE purposes).

Surface meshing versus Delaunay-type methods

The notion of a Delaunay surface mesh is in some ways confusing. The Delaunay criterion, the key to Delaunay triangulation methods, indicates that the circumdisk (circumball in three dimensions) of the mesh elements is empty. For a surface, this notion is meaningless for a number of reasons. First, such disks are not defined; second, the Delaunay criterion (that correlates to a proximity and visibility criterion) does not include any concern about curvatures (and thus directions). Therefore, Delaunay meshing of a surface is unlikely to be suitable. Nevertheless, parametric surfaces can be meshed by means of a Delaunay method, where the Delaunay property applies in the parametric space (thus a planar region) and is necessarily an anisotropic version of the method (to handle the curvature of the true surface), as discussed in the sequel.

To be more precise, while the notion of a Delaunay surface mesh is confusing, the notion of Delaunay-conforming (or admissible) surface meshes is well founded. Such a mesh enjoys the following property: inserting, by means of a three-dimensional Delaunay algorithm, the endpoints of the given surface triangles results in a volume mesh (thus a tetrahedral mesh) where all the triangles in this surface mesh are facets of elements in the volume mesh.

3.4 Combined methods

The three main classes of automated mesh-generation methods have both advantages and drawbacks. Combining one

or more of these methods is therefore an elegant way of benefiting from the advantages of such a method while avoiding any possible drawbacks. In this respect, Delaunay-type methods have been combined with octree or advancing-front techniques and, conversely, octree or advancing-front techniques have been combined with Delaunay methods.

4 QUALITY MESHING AND ADAPTIVITY

A quality mesh, or more precisely a mesh adapted to given quality and density requirements, is introduced as an occurrence of a more general class of meshes. We first introduce the notion of a regular mesh and we show how this allows for the definition of an adapted mesh.

4.1 Regular mesh

Let Ω be a closed bounded domain in \mathbb{R}^2 or \mathbb{R}^3 defined by its boundary Γ . A quality simplicial mesh or a *regular* mesh of Ω is a mesh whose elements are equilateral (regular). The existence of such a mesh is not guaranteed in general. Indeed, it depends, to some degree, on the domain boundary discretization. Therefore, we will call a simplicial regular mesh the 'best' simplicial mesh that can be completed. As the issue of constructing a regular mesh for an arbitrary domain is an open problem, there exist various methods that allow the construction of 'almost' regular meshes.

In a classical context, two types of boundary discretizations can be envisaged. The first case concerns uniform discretizations where a constant step size is given. The main advantage of such a discretization is that, in principle, it is possible to complete a regular mesh. Nevertheless, this does not guarantee a good approximation of the domain boundaries for a given step size. Given a uniform discretization of a domain boundary, a regular mesh is nothing more than a mesh where the element sizes are 'equal' to the step size serving at the boundary discretization. Thus, the desired size for the elements in the mesh is known a priori at each mesh vertex in the regular domain mesh. Let us consider the case where the domain boundary is composed of several connected components and where these are discretized by means of different step sizes (as is the case for the domains encountered in computational fluid dynamics (CFD)). A regular mesh of such a domain is a mesh where the element sizes in the neighborhood of each component are close to the step size of this component. As for the element sizes elsewhere in the domain, they must be close to the step sizes of the discretization of the boundaries situated in some neighborhood.

The second type of discretization concerns the so-called 'geometric' discretizations that are adapted to the boundary geometries. In this case, it may be proved that the discretization step size must be locally proportional to the minimum radius of curvature of the boundary. The drawback of this type of discretization is that a rather wide variation in the discretization step size may result. In order to avoid this phenomenon, a smoothing technique on the discretization step size may be applied. For a geometric discretization (not uniform in general) of the domain boundaries, it is tedious to find a priori what element sizes make the mesh regular. Obviously, a regular mesh is one where the element sizes are almost constant (or vary just a little). The idea is then to find among all the continuous size functions the function that leads to a minimal variation. This notion of a minimal variation is a characterization, among others, of the surfaces defined by means of harmonic functions.

4.2 From regular mesh to adaptivity

The two above types of meshes appear to be a particular occurrence of a more general mesh-generation problem that involves constructing a mesh whose element sizes conform to some given specifications. These requests define, in each point in the domain, the desired element sizes in all directions. There are two types of specifications: *isotropic* and *anisotropic*. In the first case, the size remains constant in all directions, which is the case we come across in mesh-adaptation problems (and thus the classical cases fall into this category). As for the second case, the size may vary when the direction varies. This is used for problems where the solution shows large variations in some directions. In both cases, we assume that we are given a function $h(P, \vec{d}) > 0$ defining the size h at point P in the domain following the direction \vec{d} , and the problem comes down to completing a mesh where the edge lengths conform to this function (or size map) h . Written in this way, the problem is not well posed. Indeed, if PQ stands for an edge in the desired mesh, the Euclidean length $\|PQ\|$ of PQ must satisfy, at the same time, the two antagonist relations

$$\|\vec{PQ}\| = h(P, \vec{PQ}) \quad \text{and} \quad \|\vec{PQ}\| = h(Q, \vec{QP})$$

which implies that the vertices P and Q are constructed in such a way as

$$h(P, \vec{PQ}) = h(Q, \vec{QP})$$

This is unlikely to be possible if the size map is such that $\forall P, Q \quad h(P, \vec{PQ}) \neq h(Q, \vec{QP})$. In fact, the computation of the Euclidean length of PQ does not take into account

the variation in the map h . Therefore, the mesh-construction problem must be formulated in a different way. To this end, we introduce a new definition. A mesh conforms to a size map h if all of its edges have an 'average' length equal to the average of the sizes specified along these edges. Thus, it is necessary to define this notion of average length with regard to a size map. To do this, we assume that the data of the map $h(X, \vec{d}) (\forall X, \vec{d})$ allows for the local definition of a metric tensor (or a metric, in short) $\mathcal{M}(X)$ at X , which in turn defines a new norm that takes into account the size variation related to the directions. Let $l_{\mathcal{M}(X)}(e)$ be the length of edge e computed using metric $\mathcal{M}(X)$. The average length of edge e may be defined as the average of the lengths $l_{\mathcal{M}(X)}(e)$ when X moves along e . If $l_m(e)$ denotes this length (subscript m for mean), we have

$$l_m(e) = \frac{\left| \int_X l_{\mathcal{M}(X)}(e) dX \right|}{\left| \int_X dX \right|}$$

for $e = PQ$ and $X = P + t\vec{PQ}$, we obtain

$$l_m(PQ) = \int_0^1 l_{\mathcal{M}(P+t\vec{PQ})}(PQ) dt$$

If $h_m(PQ)$ stands for the average of the sizes along PQ , we have

$$h_m(PQ) = \int_0^1 h(P + t\vec{PQ}, \vec{PQ}) dt$$

and edge PQ conforms to map h if $l_m(PQ) = h_m(PQ)$. To avoid computing $h_m(PQ)$, we redefine metric \mathcal{M} in such a way that $h_m(PQ) = 1$ holds, which implies (in general) $h(X, \vec{d}) = 1 (\forall X, \vec{d})$ and, in this case, we simply have $l_m(PQ) = 1$. In what follows, we give some remarks about the construction of metric $\mathcal{M}(X)$ starting from the size map $h(X, \vec{d}) (\forall X, \vec{d})$. The key is to find a metric $\mathcal{M}(X)$ that conforms as well as possible to the map. Let us recall that a metric $\mathcal{M}(X)$ defined at point X is the data of a symmetric positive-definite matrix also denoted by $\mathcal{M}(X)$. The geometric locus of the points Y that conform to metric $\mathcal{M}(X)$ at point X is in general an ellipsoid $\mathcal{E}(X)$ whose equation can be written as

$${}^t\vec{XY} \mathcal{M}(X) \vec{XY} = 1$$

This particular expression of the metric prescribes a desired size, which is unity in this metric. Indeed, for each point Y in $\mathcal{E}(X)$, we have

$$l_{\mathcal{M}(X)}(XY) = (\vec{XY}, \mathcal{M}(X) \vec{XY}) = 1$$

The set of points in $\mathcal{E}(X)$ is termed as the unit sphere associated with metric $\mathcal{M}(X)$. If $\mathcal{C}(X)$ denotes the geometric locus of the points conforming to the size map $h(X, \vec{d})$ ($\forall \vec{d}$) at point X , then metric $\mathcal{M}(X)$ may be characterized as the one whose unit sphere $\mathcal{E}(X)$ has a maximal volume included in $\mathcal{C}(X)$. Such a metric is called the underlying metric. In the particular case where the size map $h(X, \vec{d})$ only depends on X (isotropic case), metric $\mathcal{M}(X)$ reduces to

$$\mathcal{M}(X) = \frac{1}{h^2(X)} \mathcal{I}_d$$

where \mathcal{I}_d is the identity matrix in \mathbb{R}^d ($d = 2$ or 3). In this case, $\mathcal{E}(X)$ and $\mathcal{C}(X)$ are the sphere centered in X whose radius is $h(X)$. When $\mathcal{C}(X)$ is an ellipsoid (classical anisotropic case), metric $\mathcal{M}(X)$ is obviously the one whose associated sphere $\mathcal{E}(X)$ is similar to $\mathcal{C}(X)$. In the general case where $\mathcal{C}(X)$ is arbitrary, we may make use of optimization algorithms to find $\mathcal{E}(X)$ and thus to have metric $\mathcal{M}(X)$.

The size map $h(X, \vec{d})$ ($\forall X, \vec{d}$) is then seen as the metric map $\mathcal{M}(X)$, and a mesh conforming to this metric is a mesh where the edges have an average length unit. A mesh with this property is said to be a *unit mesh*. One could observe that this average edge length in a metric is nothing other than an edge length if we associate a Riemannian structure with the domain, this structure being defined by the metric map $\mathcal{M}(X)$. In this structure, length L_M of edge PQ is given by

$$L_M(PQ) = \int_0^1 \sqrt{\vec{P}\vec{Q} \mathcal{M}(P + t\vec{P}\vec{Q}) \vec{P}\vec{Q}} dt$$

To summarize, a mesh is said to be conformal to a given size map $h(X, \vec{d})$ ($\forall X, \vec{d}$) if it is unitary in the Riemannian structure associated with the underlying metric map.

From a practical point of view, the Riemannian structure may be defined in two ways: as a continuous or a discrete structure. The first way consists in defining the metric map \mathcal{M} analytically and, in this case, the metric map \mathcal{M} is explicitly given as a function of the position. The second way consists in defining the map by means of interpolation from the data of the map at the vertices of a mesh, termed the *background mesh*. This approach is popular in adaptive schemes where the metric map is computed in a supporting mesh using an appropriate error estimator. Several interpolation schemes can be used (Borouchaki *et al.*, 1997). If the map is defined in a continuous manner, then the desired unit mesh can be obtained in one iteration step. On the other hand, if the map is known in a discrete way, several iteration steps (or adaptive meshing steps) may be necessary.

To conclude, the question facing us here is to know whether a unit mesh conforming to a metric map is suitable for finite element purposes (for instance, in terms of convergence issues). To decide on this point, it is natural to add another criterion to qualify what appears to be a mesh suitable for computation. This criterion is related to the shape of the elements. A unit mesh conforming to a metric map may not be suitable for computational purposes. Indeed, the element shape quality largely depends on the size variation present in the metric map. To avoid this, it is only necessary to modify the metric map (Borouchaki *et al.*, 1997), in accordance with the desired size (while preserving certain properties included in the map). However, this modification in the metric map generally results in a larger number of elements, thus leading to a high computational cost. One possible compromise is to modify the metric map as a function of the 'sizes' that are suitable in the targeted computational method.

4.3 Unit meshing of a curve or a surface

Throughout this section, we propose a method that results in the construction of a unit mesh in a domain Ω in \mathbb{R}^d , $d = 2$ or 3 (the domain being defined by its boundary Σ), equipped with a given Riemannian metric \mathcal{M}_d . The method reduces to meshing Ω in such a way that the edges in this mesh are *unitary*. Bear in mind that the metric at a point P in Ω is defined by a symmetric positive-definite $d \times d$ matrix, $\mathcal{M}_d(P)$. If P is a vertex in the unit mesh of Ω and if PX is an edge with P as an endpoint, then one must have

$$\int_0^1 \sqrt{\vec{P}\vec{X} \mathcal{M}_d(P + t\vec{P}\vec{X}) \vec{P}\vec{X}} dt = 1$$

The proposed method involves two steps: the discretization of the boundary Σ of Ω by means of unit elements and the construction of a unit mesh in Ω using the above boundary discretization as input. These two steps are discussed in the following sections. Curve discretization includes two different cases based on the way the curve is defined: in a parametric form or in a discrete form. The same applies for the surface where a parametric or a discrete definition can be used.

4.3.1 Unit parametric curve discretization

We assume Σ to be defined by a mathematical (analytical) model. In two dimensions, the boundary Σ of Ω is composed of curved segments $\Gamma_i: I_i \rightarrow \mathbb{R}^2, t \mapsto \gamma_i(t)$ where I_i is a closed interval in \mathbb{R} and $\gamma_i(t)$ is a continuous function of class C^2 . The problem reduces to the discretization of a generic curved segment $\Gamma: I \rightarrow \mathbb{R}^2, t \mapsto$

$\gamma(t)$. In three dimensions, the boundary Σ is composed of parametric patches $\Sigma_i: \omega_i \rightarrow \mathbb{R}^3, (u, v) \mapsto \sigma_i(u, v)$, where ω_i is a closed bounded domain in \mathbb{R}^2 and $\sigma_i(t)$ is a continuous function of class C^2 . Similarly, the problem reduces to the discretization of a generic parametric patch $\Sigma: \omega \rightarrow \mathbb{R}^3, (u, v) \mapsto \sigma(u, v)$. In this case, the discretization includes two steps: the discretization of the boundary of Σ , which is composed of curved segments in \mathbb{R}^3 and that of Σ starting from the discretization of its boundary. Discretizing a curved segment in \mathbb{R}^2 is a particular case of the general problem of discretizing a curved segment in \mathbb{R}^3 . In what follows, we describe how to discretize such segments, then we show that discretizing a parametric patch reduces to constructing a unit mesh of a domain in \mathbb{R}^2 in accordance with a metric map induced by the intrinsic properties of the patch.

Discretization of a curved segment in \mathbb{R}^3

Let $\Gamma: [a, b] \rightarrow \mathbb{R}^3, t \mapsto \gamma(t)$ and let $\gamma(t)$ be a continuous function of class C^2 . As previously seen, the length of Γ in the Riemannian structure \mathcal{M}_3 is

$$L(\Gamma) = \int_a^b \sqrt{\gamma'(t) \mathcal{M}_3(\gamma(t)) \gamma'(t)} dt$$

In order to discretize Γ by means of unit segments, we first compute the integer value n closest to $L(\Gamma)$ (thus Γ must be subdivided into n segments), then we compute the values $t_i, 1 \leq i \leq n-1$ ($t_0 = a$ and $t_n = b$) such that

$$\frac{L(\Gamma)}{n} = \int_{t_{i-1}}^{t_i} \sqrt{\gamma'(t) \mathcal{M}_3(\gamma(t)) \gamma'(t)} dt$$

Finally, the discretization of Γ consists of the straight line segments $\gamma(t_i)\gamma(t_{i+1})$.

4.3.2 Unit discrete curve (re)discretization

There are essentially two ways to define a curve: by a continuous function as explained before or, in a discrete manner, by an ordered set of sampling points. These points can, for instance, be generated by a CAD system or by a scanning device, or they can be the result of a numerical simulation in an adaptive scheme. Generally, the goal is then to obtain from this set of given points a parametric curve that should be as smooth as possible.

In some cases, the data may be 'noisy' because of measurement or computation errors, producing an interpolating curve with a rough aspect. Smoothing techniques, including various averaging schemes, can be applied to avoid this phenomenon.

At present, given a set of points $\{P_i\}_{i=0, \dots, n}$, the problem is to find a continuous function γ defined on R with

values in \mathbb{R}^d (in the two- or three-dimensional space), and increasing real numbers t_i such that $\gamma(t_i) = P_i$ for each $i = 0, \dots, n$. In practice, the solution can be chosen amongst piecewise polynomial functions called *splines*, by analogy to the draftsman's tool consisting of a thin flexible rod made of metal, plastic, or wood. Each polynomial on an interval $[t_i, t_{i+1}]$ is usually of degree 3, being defined by the location of the extremities P_i and P_{i+1} , as are the tangent vectors at these points, thus ensuring a C^1 continuity of γ . To define a tangent vector at each point P_i , several methods are proposed. In the Catmull-Rom approach, the vector is colinear with $\vec{P}_{i-1}\vec{P}_{i+1}$ (with particular conditions at both ends P_0 and P_n of the curve). In the de Boor approach, a C^2 continuity is imposed at each intermediate point P_i and a linear system is solved, yielding a smoother aspect to the whole curve. Finally, having a geometric support defined by the interpolating function γ , a new discretization of the curve can be obtained using the method described in the previous section.

4.3.3 Unit parametric surface meshing

Let Σ be a parametric surface defined by $\sigma: \Delta \rightarrow \mathbb{R}^3, (u, v) \mapsto \sigma(u, v)$, Δ being a domain of \mathbb{R}^2 , and σ being a continuous function of class C^2 . We assume that Δ is closed and bounded, as is Σ . The problem we face is to construct a mesh $\mathcal{T}_{\mathcal{M}_3}(\Sigma)$ of Σ that conforms to the metric map \mathcal{M}_3 . The idea is to construct a mesh in the parametric domain Δ , so as to obtain the desired mesh after being mapped onto the surface. To this end, we show that we can define a metric \mathcal{M}_2 in Δ such that the relation $\mathcal{T}_{\mathcal{M}_3}(\Sigma) = \sigma(\mathcal{T}_{\mathcal{M}_2}(\Delta))$ is satisfied. To do so, first, we recall the usual Euclidean length formula of a curved segment of \mathbb{R}^3 plotted on Σ , and then we extend this notion to the case where a Riemannian metric is specified in Σ .

Let Γ be a curved segment of Σ defined by a continuous function of class $C^2, \gamma(t) \in \mathbb{R}^3$, where $t \in [a, b]$. The usual Euclidean length $L_{\mathbb{R}^3}(\Gamma)$ of Γ is given by

$$L_{\mathbb{R}^3}(\Gamma) = \int_a^b \sqrt{\gamma'(t) \gamma'(t)} dt$$

As Γ is plotted on Σ , there is a function $\omega(t) \in \Omega$, where $t \in [a, b]$, such that $\gamma = \sigma \circ \omega$. We have $\gamma'(t) = \sigma'(\omega(t))\omega'(t)$, where $\sigma'(\omega(t))$ is the 3×2 matrix defined as

$$\sigma'(\omega(t)) = \begin{pmatrix} \sigma'_u(\omega(t)) & \sigma'_v(\omega(t)) \end{pmatrix}$$

Thus, we obtain

$$\gamma'(t)\gamma'(t) = \omega'(t)' \sigma'(\omega(t)) \sigma'(\omega(t)) \omega'(t)$$

but we have

$${}^t\sigma'(\omega(t))\sigma'(\omega(t)) = \begin{pmatrix} {}^t\sigma'_s(\omega(t)) \\ {}^t\sigma'_e(\omega(t)) \end{pmatrix} \begin{pmatrix} \sigma'_s(\omega(t)) & \sigma'_e(\omega(t)) \end{pmatrix}$$

or

$${}^t\sigma'(\omega(t))\sigma'(\omega(t)) = \mathcal{M}_\sigma(\omega(t))$$

where \mathcal{M}_σ is a 2×2 matrix, which characterizes the local intrinsic metric of Σ at point $\gamma(t)$ and is defined as

$$\mathcal{M}_\sigma(\omega(t)) = \begin{pmatrix} {}^t\sigma'_s(\omega(t))\sigma'_s(\omega(t)) & {}^t\sigma'_s(\omega(t))\sigma'_e(\omega(t)) \\ {}^t\sigma'_e(\omega(t))\sigma'_s(\omega(t)) & {}^t\sigma'_e(\omega(t))\sigma'_e(\omega(t)) \end{pmatrix}$$

We can deduce that

$$L_{\mathcal{B}_3}(\Gamma) = \int_a^b \sqrt{{}^t\omega'(t)\mathcal{M}_\sigma(\omega(t))\omega'(t)} dt$$

The above formula has an interesting interpretation. The Euclidean length of the curved segment $\omega(t)$ plotted in Ω depends on the Euclidean norm of $\omega'(t)$, while the Euclidean length of the curved segment $\gamma(t) = \sigma(\omega(t))$ (image of $\omega(t)$ on Σ) depends on the 'Riemannian norm' of $\omega'(t)$ with respect to the local intrinsic metric of Σ .

In particular (depending on the particular reason for generating a mesh), if $\omega(t)$ is a line segment AB of Ω , we have $\omega(t) = A + t\vec{AB}$; thus $\omega'(t) = \vec{AB}$, and

$$L_{\mathcal{B}_3}(\sigma(AB)) = \int_0^1 \sqrt{{}^t\vec{AB}\mathcal{M}_\sigma(A + t\vec{AB})\vec{AB}} dt$$

The above formula allows us to compute the length of the curved segment on Σ , which is the mapping by σ of an edge plotted on Ω . If the new metric \mathcal{M}_σ is given in Ω , we have

$$L_{\mathcal{M}_\sigma}(AB) = \int_0^1 \sqrt{{}^t\vec{AB}\mathcal{M}_\sigma(A + t\vec{AB})\vec{AB}} dt = L_{\mathcal{B}_3}(\sigma(AB))$$

Let us consider the case where a Riemannian metric \mathcal{M}_3 of \mathbb{R}^3 is specified on Σ . In this case, the Riemannian length $L_{\mathcal{M}_3}(\Gamma)$ of Γ is given by

$$L_{\mathcal{M}_3}(\Gamma) = \int_a^b \sqrt{{}^t\gamma'(t)\mathcal{M}_3(\gamma(t))\gamma'(t)} dt$$

which can be written as

$$L_{\mathcal{M}_3}(\Gamma) = \int_a^b \sqrt{{}^t\omega'(t)\tilde{\mathcal{M}}_2(\omega(t))\omega'(t)} dt$$

with

$$\tilde{\mathcal{M}}_2(\omega(t)) = {}^t\sigma'(\omega(t))\mathcal{M}_3(\gamma(t))\sigma'(\omega(t))$$

Thus, if $\omega(t)$ is a line segment AB of Ω , we obtain

$$L_{\mathcal{M}_3}(\sigma(AB)) = \int_0^1 \sqrt{{}^t\vec{AB}\tilde{\mathcal{M}}_2(A + t\vec{AB})\vec{AB}} dt$$

The above formula allows us to compute the generalized length of the curved segment on Σ , which is the mapping by σ of an edge plotted on Ω . Let us define the new metric $\tilde{\mathcal{M}}_2$ in Δ . We have

$$L_{\tilde{\mathcal{M}}_2}(AB) = \int_0^1 \sqrt{{}^t\vec{AB}\tilde{\mathcal{M}}_2(A + t\vec{AB})\vec{AB}} dt$$

and thus we obtain

$$L_{\mathcal{M}_3}(\sigma(AB)) = L_{\tilde{\mathcal{M}}_2}(AB)$$

To return to the problem of mesh generation itself, the last equation shows that the mapping of the mesh conforming to the metric $\tilde{\mathcal{M}}_2$ of Δ onto the surface Σ gives the mesh conforming to the specified metric \mathcal{M}_3 of Σ .

4.3.4 Unit discrete surface meshing

The problem of meshing a discrete surface (a surface defined by a triangulation) is closely related to the problem of defining a suitable metric map \mathcal{M}_3 on the surface to control mesh modifications. The idea is to extract the intrinsic properties of the underlying surface from the initial surface triangulation.

Formally speaking, the problem is to construct a geometric surface mesh $\mathcal{T}_{\mathcal{M}_3}$ from an initial reference mesh \mathcal{T}_{ref} . At first, mesh \mathcal{T}_{ref} is simplified and optimized in accordance with a Hausdorff distance δ , resulting in a geometric reference mesh $\mathcal{T}_{ref,\delta}$. This first stage aims at removing extra vertices (i.e. those that do not contribute explicitly to the surface definition) and at denoising the original data. This procedure consists in removing the mesh vertices iteratively, provided two conditions are satisfied: an approximation criterion (related to the Hausdorff distance between the current mesh and the original one) and a regularity criterion (related to the local deviation of the mesh edges from the local tangent planes). Then, a geometric support, piecewise C^1 continuous, is defined on mesh $\mathcal{T}_{ref,\delta}$ so as to define a 'smooth' representation of the surface Σ . The aim of this support is to supply the location of the closest point onto the surface, given a point on a triangular facet. This support will be used to insert a new vertex in the current mesh.

Curvature evaluation

In order to build the metric map \mathcal{M}_3 , we need to evaluate the intrinsic properties of the surface. This requires computing the principal curvatures and principal directions at the

vertices of mesh $\mathcal{T}_{ref,\delta}$. To this end, the surface at a point P is locally approached by a quadric surface, based on a least square fit of adjacent mesh vertices. The local frame at P is computed on the basis of the normal to a discrete approximation of the normal to the surface at P . To find the coefficients of the quadric, we consider all vertices P_i of the ball of P and assume that the surface fits, at best, these points. Solving this system is equivalent of minimizing the sum

$$\min \sum_i (ax_i^2 + bx_iy_i + cy_i^2 - z_i)^2$$

which corresponds to minimizing the square of the norm of the distance to the quadric surface. Knowing the coefficients a , b , and c , it becomes easy to find the local curvatures at a point $P = (0, 0, 0)$ in the local frame:

$$\begin{aligned} E &= 1 + (2au + bv)^2 = 1, & L &= 2a \\ F &= (2au + bv)(bu + 2cv) = 0, & M &= b \\ G &= 1 + (bu + 2cv)^2 = 1, & N &= 2c \end{aligned}$$

The analysis of the variations of the normal curvature function

$$\kappa_n(\vec{\tau}) = \frac{\Phi_1(\vec{\tau})}{\Phi_2(\vec{\tau})}$$

leads to resolving a second-order equation that admits (in principle) two distinct solutions, two pairs (λ_1, κ_1) , (λ_2, κ_2) . The extrema values κ_1 and κ_2 of κ_n (i.e. the roots of the equation) are the principal curvatures of the surface at P . Considering the second-order equation

$$\kappa^2 - (\kappa_1 + \kappa_2)\kappa + \kappa_1\kappa_2 = 0$$

yields

$$\begin{aligned} \kappa_1\kappa_2 &= \frac{LN - M^2}{EG - F^2} = 4ac - b^2, \\ \kappa_1 + \kappa_2 &= \frac{NE - 2MF + LG}{EG - F^2} = 2(a + c) \end{aligned} \quad (1)$$

where $K = \kappa_1\kappa_2$ is the Gaussian curvature and $H = 1/2(\kappa_1 + \kappa_2)$ is the mean curvature of the surface at P . Solving these equations allows us to find the extrema values κ_1 and κ_2 at P

$$\kappa_i = \frac{2(a + c) \pm \sqrt{\Delta}}{2}$$

where $\Delta = (2(a + c))^2 - 4(4ac - b^2)$. A similar analysis leads to finding the principal directions at a point P on the surface.

Metric definition

A geometric metric map $\mathcal{M}_3(P)$ can be defined at any mesh vertex P so as to locally bind the gap between the mesh edges and the surface by any given threshold value ε . A matrix of the form

$$\mathcal{M}_3(P)_{\rho_1, \rho_2} = {}^t\mathcal{D}(P) \begin{pmatrix} \frac{1}{\alpha^2 \rho_1^2(P)} & 0 & 0 \\ 0 & \frac{1}{\beta^2 \rho_2^2(P)} & 0 \\ 0 & 0 & \lambda \end{pmatrix} \mathcal{D}(P)$$

where $\mathcal{D}(P)$ corresponds to the principal directions at P , $\rho_1 = 1/\kappa_1$, $\rho_2 = 1/\kappa_2$ are the main radii of curvature, α and β are appropriate coefficients and $\lambda \in \mathbb{R}$ provides an anisotropic (curvature-based) control of the geometry. This discrete metric prescribes mesh sizes as well as element stretching directions at mesh vertices. The local size is proportional to the principal radii of curvature, the coefficient of proportionality being related to the largest allowable deviation gap between the mesh elements and the surface geometry (Frey and Borouchaki, 1998). For instance, setting constant gap values comes down to fixing

$$\alpha = 2\sqrt{\varepsilon(2 - \varepsilon)} \quad \text{together with} \quad \beta = 2\sqrt{\varepsilon \frac{\rho_1}{\rho_2} \left(2 - \varepsilon \frac{\rho_1}{\rho_2}\right)}$$

As the size may change rapidly from one vertex to another, the mesh gradation may not be bounded locally. To overcome this problem, size map \mathcal{M}_3 is modified using a size-correction procedure (Borouchaki et al., 1997).

Surface remeshing

Having defined an adequate metric \mathcal{M}_3 at any mesh vertex, the discrete surface-meshing problem consists in constructing a unit mesh with respect to metric map \mathcal{M}_3 . To this end, local mesh modifications are applied on the basis of the edge length analysis. The optimization consists in collapsing the short edges and splitting the large edges on the basis of their relative length. Geometric measures are used to control the deviation between the mesh elements and the surface geometry, as well as the element shape quality (Frey and Borouchaki, 1998). A point relocation procedure is also used in order to improve the element shape quality.

4.4 Unit volume meshing

The global scheme for unit mesh generation is well known: a coarse mesh (without internal points) of the domain is constructed using a classical Delaunay method, and then this mesh is enriched by adding the field points before being optimized. The field points are defined in an iterative

manner. At each iteration step, these points are created using a method for edge saturation or an advancing-front method. Then, they are inserted in the current mesh using the constrained Delaunay kernel (George and Borouchaki, 1998) in a Riemannian context. This process is repeated as long as the current mesh is being modified. On the basis of this method, the field points are necessarily well located with respect to the mesh entities that have already been constructed. Similarly, the Delaunay kernel results in almost optimal connections (an optimal connection is found when regular elements exist) from point to point. In what follows, we give some details about the field point placement strategies and we discuss the generalized constrained Delaunay kernel.

4.4.1 Point placement strategies

Here, we propose two methods that allow the field points to be defined. The advantage of the first is its simplicity and low cost. On the other hand, the second method generally results in better quality meshes.

Edge saturation

At each iteration step, the field points are defined in such a way as to subdivide the edges in the current mesh by unit length segments. A field point is retained if it is not too close (i.e. at a distance less than 1) to a point already existing in the mesh or previously retained for insertion.

Advancing-front strategy

At each iteration step, a set of facets (edges in two dimensions) in the current mesh, thus a front, is retained and the field points are created from this front so as to form unit elements (elements with unit edges). A facet in the current mesh is considered as a front entity if it separates a unit element from a nonunit element (nu for short). Particular care is necessary if one wants to properly cover (meaning that the number of points thus defined is adequate) the whole domain. For instance, a nonunit element may be seen as a unit element at some time. The optimal point related to a front facet is defined in the same side as where the associated nonunit element lies. This is the point that results in a unit element after being connected to the front entity. Let f_i be a front facet at iteration step i , which separates the unit element $K_i^u = (f_i, P_i)$, where P_i is the vertex in K_i^u other than the endpoints of f_i from the nonunit element $K_i^n = (f_i, P_i)$, where P_i is the vertex in K_i^n other than the endpoints of f_i . The optimal point P_i^* with respect to f_i is constructed in such a way that element (f_i, P_i^*) is unitary. If point P_i^* is at a distance less than one from P_i , then the nonunit element K_i^n is considered as a unit element in the next iteration step $i+1$. This means that the nonunit

element K_i^n is constructed and nothing better can be done. A variation leads to considering all the nonunit elements that have a front facet at iteration step i as unit elements at iteration step $i+1$. Obviously, these elements will be considered (when defining the front at iteration step $i+1$) if they are not affected by any point insertion at iteration step i . As above, the optimal point for a front facet is created if it falls inside the domain and if it is not too close to an existing point.

4.4.2 Point insertion

In a classical problem, the constrained Delaunay kernel based on cavity remeshing is written as (cf. Watson, 1981) $T = T - C(P) + B(P)$, where $C(P)$ is the cavity associated with point P and $B(P)$ is the remeshing of $C(P)$ based on P (T being the current mesh). The cavity is constructed following a constrained proximity criterion given by

$$\{K, K' \in T, P \in \text{Ball}(K) \text{ and } P \text{ visible from each vertex in } K\}$$

where $\text{Ball}(K)$ is the opened circumball of K .

An extension of this approach involves redefining the cavity $C(P)$ in a Riemannian context (George and Borouchaki, 1998). To this end, we first introduce the Delaunay measure α_{M_d} associated with pair (P, K) , with respect to a metric M_d :

$$\alpha_{M_d}(P, K) = \left[\frac{d(O_K, P)}{r_K} \right]_{M_d}$$

where O_K (resp. r_K) is the center (resp. radius) of the circumsphere of K and $[*]_{M_d}$ indicates that the quantity $*$ is measured in the Euclidean space characterized by metric M_d . The usual proximity criterion, $P \in \text{Ball}(K)$, is expressed by $\alpha_{T_d}(P, K) < 1$, where T_d is the identity metric. Cavity $C(P)$ is then redefined by $C(P) = C_1(P) \cup C_2(P)$ with

$$C_1(P) = \{K, K' \in T, K \text{ including } P\}$$

$$C_2(P) = \{K, K' \in T, \exists K' \in C(P), K \text{ adjacent to } K'\}$$

$$\alpha_{M_d}(P) = \sum_V \alpha_{M_d}(V)(P, K) < d+2,$$

$$V \text{ vertex of } K, P \text{ visible by the vertices of } K\}$$

Hence, region $C(P)$ is completed by adjacency starting from the elements in $C_1(P)$. After this definition, and in two dimensions, the generalized cavity is star-shaped with respect to P and the $B(P)$ is valid. In three dimensions, a possible correction (George and Hermeline, 1992) of the cavity based on element removal can ensure this property.

4.4.3 Optimization processes

The proposed method results in a unit mesh of domain Ω . Nevertheless, the mesh quality can be improved by means of two optimization processes, one made up of topological modifications, the other consisting of geometrical modifications. The first mainly involves applying a number of facet flips, while the other consists of node repositioning. In fact, we assume that the mesh of domain Ω (before being optimized) has more or less the right number of internal vertices. These optimization procedures do not modify the number of internal nodes but, in particular, enhance the quality of the nonunit elements created when the field points have been generated. The scheme consists of iterative facet flips and node repositioning. In what follows, we recall the notion of edge quality and element quality and we discuss these optimization tools.

Edge length quality

Let AB be a mesh edge. The length quality Q_l of AB in the Riemannian metric M_d may be defined as

$$Q_l(AB) = \begin{cases} L_{M_d}(AB) & \text{if } L_{M_d}(AB) \leq 1 \\ \frac{1}{L_{M_d}(AB)} & \text{if } L_{M_d}(AB) > 1 \end{cases}$$

With this measure, $0 \leq Q_l(AB) \leq 1$ holds and a unit edge has a length quality with a value of 1. This quality measure about the edge lengths shows how the mesh conforms to the specified Riemannian metric M_d .

The edge length quality of a mesh T is defined by

$$Q_l(T) = \left(\frac{1}{|T|} \sum_{e \in T} Q_l(e), \min_{e \in T} Q_l(e) \right)$$

where e stands for an edge in mesh T and $|T|$ is the number of such edges. The two quantities in the formula measure respectively the average and the minimum of the length qualities of the mesh edges.

Element shape quality

Let K be a mesh element. In the classical Euclidean space, a popular measure for the shape quality of K is (Lo, 1991)

$$Q_f(K) = c \frac{V(K)}{\sum_{e(K)} l^2(e(K))}$$

where $V(K)$ denotes the volume of K , $e(K)$ being the edges in K and c the scaling coefficient such that the quality of a regular element has the value of 1. With this definition, we have $0 \leq Q_f(K) \leq 1$ and a nicely shaped element has a

quality close to 1, while an ill-shaped element has a quality close to 0.

In a Riemannian space, the quality of element K can be defined by

$$Q_f(K) = \min_{1 \leq i \leq d+1} Q_f^i(K)$$

where $Q_f^i(K)$ is the element quality in the Euclidean space associated with the metric M_{d_i} corresponding to the vertex number i in K .

To measure the quality $Q_f^i(K)$, it is simply necessary to transform the Euclidean space related to the metric specified at vertex i of K into the usual Euclidean space and to consider the quality value of element K^i associated with K ; in other words

$$Q_f^i(K) = Q_f(K^i)$$

It is easy to show that

$$Q_f^i(K) = c \frac{\sqrt{\text{Det}(M_{d_i}^i)} V(K)}{\sum_{e(K)} l_{M_{d_i}^i}^2(e(K))}$$

Similarly, the shape quality of the elements in mesh T is defined by

$$Q_f(T) = \left(\frac{1}{|T|} \sum_{K \in T} Q_f(K), \min_{K \in T} Q_f(K) \right)$$

where K stands for an element in mesh T . The two quantities in the formula measure respectively the average and the minimum shape qualities of the mesh elements.

Facet flip

Facet flip only affects the mesh topology. It has proved to be very efficient for shape quality improvement. This technique results in the removal of a facet of arbitrary dimensionality whenever this is possible. Let f be a facet of arbitrary dimensionality in the mesh. We use the term *shell* (cf. George and Borouchaki, 1998) for f , the set of elements sharing f . Flipping f involves constructing a triangulation of the hull of the shell of f , where f is not a mesh entity. The quality of a shell is that of its worst element. The flip is then processed if the quality of the new triangulation is better than that of the initial shell.

When a Riemannian metric must be followed, it is necessary to sort these facet flips, while this is not strictly necessary in a classical Euclidean case. This leads to the association of the expected ratio of improvement β_f with face f by emulating a flip. Then, to optimize the mesh, an iterative process is used, which applies the flips in the

decreasing order of the above ratios. To begin with, the ratio of improvement is set at a given value $\omega > 1$, then ω is modified and decreases to 1. Such a strategy leads to flipping in first the most significative operations in terms of mesh improvement.

Point repositioning

Let P be an internal point in the mesh and let (K_i) be the set of elements with P as a vertex (i.e. the ball associated with P (cf. George and Borouchaki, 1998)). Repositioning P consists in moving P so as to enhance the quality of the worst elements in (K_i) . Two methods can be advocated for node repositioning. One based on unit length, the other on optimal elements. The first method improves the edge length quality for the elements in (K_i) , while the second improves the shape of these elements. In practice, both methods are applied to all the internal points of the mesh.

Let (P_i) be the set of vertices in (K_i) other than P . With each point P_i is associated the optimal point (P_i^*) such that

$$P_i^* P_i^* = \frac{1}{L_{M_0}(P_i P_i^*)} P_i^* P_i^*$$

for which $L_{M_0}(P_i P_i^*) = 1$ holds. Repositioning P consists of moving point P step by step toward the centroid of the points (P_i^*) if the quality of the worst element in (K_i) is improved. This process results in unit edge lengths for the edges that have P as one endpoint.

Let (f_i) be the facets in the elements in (K_i) , which are opposite vertex P , that is, $(K_i) = [P, f_i]$. With each facet f_i , the optimal point P_i^* is associated such that the element $K_i^* = [P_i^*, f_i]$ satisfies

$$Q_f(K_i^*) = \max_{\tilde{P}} Q_f([\tilde{P}, f_i])$$

where \tilde{P} is an arbitrary point located on the same side of f_i as P is. Similarly, repositioning P consists of moving P step by step toward the centroid of points (P_i^*) if the quality of the worst element in (K_i) is improved. This process results in optimal quality elements for the elements in the ball of P .

To find point P_i^* , we can consider the centroid of the optimal points related to the f_i 's, each of which is evaluated in the Euclidean structure related to the metric defined at a vertex of K_i .

5 ADAPTIVE FEM COMPUTATIONS

Let us consider a bounded domain Ω described by its boundary surface Γ . In the context of a numerical simulation performed on domain Ω using a mesh of Ω as

spatial support, we suggest a general scheme including an adaptive meshing loop of Ω . This scheme is made up of two distinct parts. The first part only involves the generation of an initial mesh of the computational domain Ω . The second part concerns an adaptation loop including the computation, the a posteriori error estimation, and the generation of adapted meshes. The method advocated in what follows is an h -method, where the parameter of adaptation is h , for example, the size (or the directional sizes) of the mesh elements. Other adaptation methods exist, including p -methods, hp -methods, and hierarchical methods, as well as some others (such as local refinement) (see Chapter 4, this Volume, Chapter 2, Volume 2).

5.1 Initial mesh of Ω

The problem consists of constructing a mesh of Ω from an initial reference mesh $T_{ref}(\Gamma)$ of its boundary Γ and from a metric map $M_0(\Gamma)$, indicating the desired element sizes on the surface. To construct this initial mesh $T_0(\Omega)$, we proceed in several steps:

- The initial reference mesh $T_{ref}(\Gamma)$ is simplified and optimized in a given Hausdorff envelope, so as to obtain a geometric reference mesh $T_{ref,g}(\Gamma)$ of Γ
- A geometric 'smooth' support (piecewise G^1 continuous) is defined on mesh $T_{ref,g}(\Gamma)$, so as to obtain a smooth geometric representation of boundary Γ
- Metric map $M_0(\Gamma)$, supplied on mesh $T_{ref,g}(\Gamma)$ is then rectified so as to be compatible with the surface geometry
- The rectified map $M_0(\Gamma)$ is again modified to account for the desired mesh gradation
- Mesh $T_{ref,g}(\Gamma)$ is adapted in terms of element sizes to the modified map $M_0(\Gamma)$ so as to obtain the initial computational mesh $T_0(\Gamma)$
- Volume mesh $T_0(\Omega)$ is generated from mesh $T_0(\Gamma)$ associated with the metric map $M_0(\Gamma)$.

This schematic flow is illustrated in Figure 1, where one can see the data flowchart related to the input and the output of the various procedures involved in the entire process.

5.2 General diagram of an adaptive computation

The adaptation loop aims to capture and to refine the physical solution of the numerical simulation performed on domain Ω . In general, a computation performed on mesh $T_0(\Omega)$ does not allow a satisfactory solution to be obtained. Hence, we suggest the following iterative adaptation scheme, in which at each iteration step i

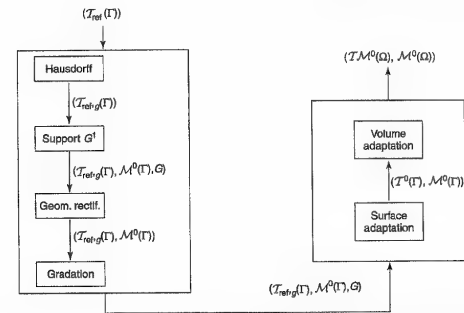


Figure 1. The various stages of the construction of the initial (volume) mesh of Ω : from the initial mesh $T_{ref}(\Gamma)$ of the boundary to the computational mesh $T_0(\Omega)$ and the associated metric map $M_0(\Omega)$.

- a computation is performed on mesh $T_i(\Omega)$, leading to the solutions field $S_i(\Omega)$;
- solution $S_i(\Omega)$ is analyzed using an adequate error estimator and a metric map $M_i(\Omega)$ is deduced, prescribing the element sizes in order to obtain a subsequent solution at a given accuracy;
- the metric map $M_i(\Gamma)$ restricted to surface Γ is rectified with regard to the surface geometry;
- the (partially) rectified metric map $M_i(\Omega)$ is modified to take into account the specified mesh gradation;
- surface mesh $T_i(\Gamma)$ is adapted in terms of element sizes to the metric modified map $M_i(\Gamma)$ to obtain mesh $T_{i+1}(\Gamma)$;
- volume mesh $T_{i+1}(\Omega)$ adapted to the metric field $M_i(\Omega)$ is generated from mesh $T_i(\Gamma)$ associated with its metric map $M_i(\Gamma)$;
- solution $S_i(\Omega)$ associated with mesh $T_i(\Omega)$ is interpolated on mesh $T_{i+1}(\Omega)$.

These various stages are illustrated in Figure 2.

5.3 Error estimates

There exist several error estimate strategies that make it possible to control a posteriori the error with the solution computed on a finite element mesh (Fortin, 2000). These estimators can be used to control the mesh by means of an h -adaptation method in such a way that the resulting solution is of a given accuracy.

Among these estimators are those that are based on the interpolation error (and, therefore, of a purely geometric nature as the operator itself is not considered). This class of estimators has been studied by various authors (Babuska and Aziz, 1976; D'Azevedo and Simpson, 1991; Rippa, 1992; Berzins, 1999). Nevertheless, most of these papers require a parameter, h , the element size, to be small or to vanish, and therefore are asymptotic results. The estimator then makes use of appropriate Taylor expansions and provides information about the desired size, h . However, as this size is not necessarily small, we propose a new approach that, while close, does not necessitate any peculiar assumption about this parameter, and, therefore, is likely to be more justified. Our approach is, to some extent, close to solutions used in a different topic, the construction of meshes in parameterized patches (see Sheng and Hirsch, 1992; Anglada *et al.*, 1999; Borouchaki *et al.*, 2001; among others).

5.3.1 Problem statement and state of the art

Let Ω be a domain in R^d (where $d = 1, 2$, or 3) and let T be a simplicial mesh of Ω , where the simplices are linear, P^1 , or quadratic, P^2 , elements. We assume that we have in hand the solution, denoted by u (to meet the classical notations in error estimate literature), of a finite element computation previously done in Ω using T as a mesh, this solution being scalar values denoted by u_T . Let u be the exact solution; the problem first involves computing the gap $e_T = u - u_T$ from u to u_T , which represents the underlying error due to the

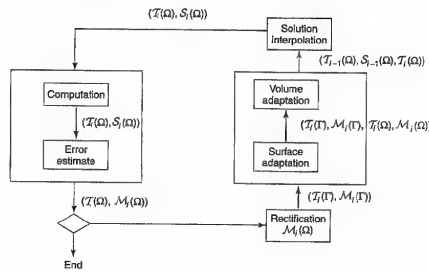


Figure 2. The various stages of the general adaptation scheme.

finite element approximation. Then, it involves constructing a new mesh T' such that the estimated gap from u to the solution u_T , as obtained using a new computation, is bounded by a given threshold. Specific issues include

- how to evaluate gap e_T from u and u_T ;
- how to use this information to construct a new mesh such that the corresponding gap is bounded in the desired range.

The finite element solution u_T is not interpolant (e.g. the solution at the nodes of T is not coincident with the exact value u at these nodes). Moreover, for any element in the mesh, it is not possible to guarantee that u_T is coincident with the exact value of u at at least one point in this element. It is therefore tedious to explicitly access the gap of e_T . However, the direct analysis of this gap has been studied in a number of works (Verfürth, 1996). But, in general, it remains an open problem. Therefore, other nondirect approaches have been proposed to quantify or bound this gap. Let \tilde{u}_T be the interpolate of u over mesh T (here is a linear or a quadratic piecewise function after the degree of the elements in T) and let \tilde{e}_T be the gap $u - \tilde{u}_T$ from u to \tilde{u}_T , the so-called *interpolation error* about u for T . To obtain the range of gap e_T , we assume the following relation to be satisfied:

$$||e_T|| \leq C ||\tilde{e}_T||$$

where $||\cdot||$ stands for a norm and C is a constant independent of T . In other words, we assume the finite element error to be majored by the interpolation error. This allows us to simplify the initial problem by considering the following problem: given \tilde{u}_T , the interpolation of u over mesh T ,

how can we construct another mesh T' where the interpolation error is bounded by a given threshold? As \tilde{u}_T can be seen as a discrete representation of u , the problem reduces to finding a characterization of the meshes where this interpolation error is bounded. This topic has been addressed in various papers, such as Berzins (1999) and, in most of them, using a 'measure' of the interpolation error makes it possible to find some constraints to which the mesh elements must conform. In the context of mesh-adaptation methods, we are mainly interested in h -methods or adaptation in size where these constraints are expressed in terms of element sizes. Some classical measures of this error, classified into two categories, continuous or discrete, together with the corresponding constraints about the mesh elements are discussed in Berzins (1999). It turns out that the discrete approach is more convenient from the point of view of mesh adaptation.

In this chapter, we first propose a majoration of the interpolation error in two dimensions that extends to arbitrary dimensions (and is close to a work by Anglada *et al.* (1999)). Then, we show how this majoration can be used to adapt the mesh. After which, we introduce a new measure for quantifying the interpolation error, which depends on the local deformation of the Cartesian surface associated with the solution. We demonstrate how this measure makes it possible to control the interpolation error in H^1 norm and, therefore, is likely to be more appropriate.

5.3.2 A bound in two dimensions

We consider a mesh made up of piecewise linear triangles. Let K be a mesh element, and let u be the mapping from R^2 to R , assumed to be sufficiently smooth. We note by

$\Pi_K u$ the linear interpolant of u over K , and we assume u and $\Pi_K u$ to be coincident at the vertices of K . The aim is to bound $|(u - \Pi_K u)(x)|$ for x in K .

An isotropic bound

In the reference George (2001) it is shown that

$$e = ||(u - \Pi_K u)(x)||_{\infty} \leq \frac{2}{9} L^2 M \quad (2)$$

where L is the longest edge in K and M is given by $\langle H_u$ being the Hessian of u

$$M = \max_{x \in K} \left(\max_{||\vec{v}||=1} |(\vec{v}, H_u(x) \vec{v})| \right)$$

After (2), a gap in the range of ε for e implies that L is such that

$$L^2 \leq \frac{9\varepsilon}{2M} \quad (3)$$

while observing that if M is zero, any value of L is convenient. Therefore, after the value of h_{\max} , the diameter of element K , is compared with L , one can know whether triangle K is suitable, too small, or too large. Therefore, a size adaptation can be envisaged if necessary.

However, in practice, the difficulty is to evaluate M using $\Pi_K u$ by means of

- approaching $\nabla_x(a)$, $\nabla_x(b)$, and $\nabla_x(c)$,
- approaching $H_x(a)$, $H_x(b)$, and $H_x(c)$,
- approaching M by the largest eigenvalue of matrices $|H_x(a)|$, $|H_x(b)|$, and $|H_x(c)|$, where $|H_x|$ is constructed from H_x after being made positive definite.

Similarly, vectors $\nabla_x(a)$, $\nabla_x(b)$, and $\nabla_x(c)$ and matrices $H_x(a)$, $H_x(b)$, and $H_x(c)$ can be approximated using generalized finite differences (a variation of the Green formula (Raviart and Thomas, 1988)) of function $\Pi_K u$. However, the Taylor expansion about a , b , and c can be used. To this end, we assume $K = [a, b, c]$ to be an element inside the mesh and we denote by (a_i) the set of the mesh vertices adjacent to a (thus including b and c). Writing the Taylor expansion with respect to a in each a_i yields the following overdetermined system

$$\{u(a_i) \approx u(a) + (\vec{a} \vec{d}_i, \nabla_x(a))\}$$

which is equivalent to the minimization problem

$$\min_{\vec{z}} \sum_i w_i^2 ((\vec{a} \vec{d}_i, \vec{z}) - u(a_i) + u(a))^2$$

where w_i is a weight, measuring the influence of equation i in the system. The solution of this problem is that of the linear system $'AWA \vec{z} = W\beta$, where A is the matrix made up of the row vectors $(\vec{a} \vec{d}_i)$, W being the diagonal matrix related to the weights (w_i^2) , and β being the vector with components $(u(a_i) - u(a))$. Similarly, vector $\nabla_x(a)$ being known, Hessian $H_x(a)$ can be computed by the linear system

$$\{u(a_i) \approx u(a) + (\vec{a} \vec{d}_i, \nabla_x(a)) + \frac{1}{2} (\vec{a} \vec{d}_i, H_x(a) \vec{a} \vec{d}_i)\}$$

which is equivalent to a similar minimization problem. Vectors $\nabla_x(b)$ and $\nabla_x(c)$, as well as matrices $H_x(b)$ and $H_x(c)$, are obtained in the same way.

Owing to the isotropic nature of this result, using this kind of estimator may lead to an unnecessarily fine mesh. In fact, M is the largest eigenvalue of the family of operators $|H_x|$ (in all points in K) and therefore L is the size corresponding to this value. As a consequence, an anisotropic phenomenon will be considered in an isotropic way while imposing, in all directions, a size equal to the smallest size related to the eigenvalues.

An anisotropic bound

We assume that vertex a of K is the site of x (i.e. x is closer to a than to b and c) to be the point where a maximal gap occurs. We also assume x to be in K (and not in one edge of K , which leads to a similar result). Then we note a' , the point of intersection of the line supporting ax with the edge opposite a , for example, edge bc in K . We expand e in a from x by means of the Taylor expansion with integral

$$\begin{aligned} e(a) &= (u - \Pi_K u)(a) = (u - \Pi_K u)(x) \\ &+ (\vec{x} \vec{a}, \nabla_x(u - \Pi_K u)(x)) \\ &+ \int_0^1 (1-t) (\vec{x} \vec{a}, H_x(x + t\vec{x} \vec{a}) \vec{x} \vec{a}) dt \end{aligned}$$

As a is the site of x , the scalar value λ such that $\vec{x} \vec{a} = \lambda \vec{a} \vec{a}'$ is smaller than $2/3$; therefore

$$\begin{aligned} |e(x)| &= \left| \int_0^1 (1-t) \lambda^2 (\vec{a} \vec{a}', H_x(a + t\vec{x} \vec{a}) \vec{a} \vec{a}') dt \right| \\ &\leq \frac{4}{9} \int_0^1 (1-t) (\vec{a} \vec{a}', H_x(a + t\vec{x} \vec{a}) \vec{a} \vec{a}') dt \end{aligned}$$

which yields

$$|e(x)| \leq \frac{2}{9} \max_{y \in K} |(\vec{a} \vec{a}', H_x(y) \vec{a} \vec{a}')| \quad (4)$$

Remark. In an arbitrary dimension, for example, d , the constant $2/9$ should be $(1/2(d/d+1)^2)$.

After (4), imposing a gap of ε for e leads for triangle $K = [a, b, c]$

$$\max_{y \in K} |\langle \vec{aa'}, H_a(y) \vec{aa'} \rangle| \leq \frac{9}{2} \varepsilon$$

where a' is defined as above. This inequality is satisfied if

$$\max_{y \in K} \langle \vec{aa'}, |H_a(y)| \vec{aa'} \rangle \leq \frac{9}{2} \varepsilon$$

Let $\mathcal{M}(a)$ be the symmetric positive-definite matrix such that

$$\max_{y \in K} \langle \vec{aa'}, |H_a(y)| \vec{aa'} \rangle \leq \langle \vec{aa'}, \mathcal{M}(a) \vec{aa'} \rangle$$

for all z in K and such that the region (e.g. bounded by the corresponding ellipse) defined by

$$\langle \vec{aa'}, \mathcal{M}(a) \vec{aa'} \rangle \leq \frac{9}{2} \varepsilon$$

has a minimal surface. Then $\mathcal{M}(a)$ results in a size constraint or again a metric in a that varies after the directions.

In these equations, we have assumed the site of point x where gap e is maximal to be a . As x is not known, the sites b and c must be taken into account in turn. This leads to the system

$$\begin{cases} \max_{y \in K} \langle \vec{aa'}, |H_a(y)| \vec{aa'} \rangle \leq \frac{9}{2} \varepsilon \\ \max_{y \in K} \langle \vec{bb'}, |H_b(y)| \vec{bb'} \rangle \leq \frac{9}{2} \varepsilon \\ \max_{y \in K} \langle \vec{cc'}, |H_c(y)| \vec{cc'} \rangle \leq \frac{9}{2} \varepsilon \end{cases}$$

where b' and c' are points respectively in edges ac and ab . As above, we can define the metrics $\mathcal{M}(a)$, $\mathcal{M}(b)$, and $\mathcal{M}(c)$ in a , b , and respectively c such that

$$\begin{cases} \max_{y \in K} \langle \vec{aa'}, |H_a(y)| \vec{aa'} \rangle \leq \langle \vec{aa'}, \mathcal{M}(a) \vec{aa'} \rangle \leq \frac{9}{2} \varepsilon \\ \max_{y \in K} \langle \vec{bb'}, |H_b(y)| \vec{bb'} \rangle \leq \langle \vec{bb'}, \mathcal{M}(b) \vec{bb'} \rangle \leq \frac{9}{2} \varepsilon \\ \max_{y \in K} \langle \vec{cc'}, |H_c(y)| \vec{cc'} \rangle \leq \langle \vec{cc'}, \mathcal{M}(c) \vec{cc'} \rangle \leq \frac{9}{2} \varepsilon \end{cases}$$

Therefore, for triangle K , a metric $\mathcal{M}(K)$ can be constructed in such a way as, on the one hand, equations

$$\begin{cases} \langle \vec{v}, \mathcal{M}(K) \vec{v} \rangle \leq \langle \vec{v}, \mathcal{M}(a) \vec{v} \rangle \\ \langle \vec{v}, \mathcal{M}(K) \vec{v} \rangle \leq \langle \vec{v}, \mathcal{M}(b) \vec{v} \rangle \\ \langle \vec{v}, \mathcal{M}(K) \vec{v} \rangle \leq \langle \vec{v}, \mathcal{M}(c) \vec{v} \rangle \end{cases}$$

are satisfied for any vector \vec{v} , and, on the other hand, the surface $\langle \vec{v}, \mathcal{M}(K) \vec{v} \rangle = 1$ is maximal. In other words, the metric in K is the largest size constraint along all the directions satisfying the size constraints at the vertices a , b , and c .

Actually, for the sake of simplicity, we consider $\mathcal{M}(a) = |H_a(a)|$, $\mathcal{M}(b) = |H_b(b)|$, and $\mathcal{M}(c) = |H_c(c)|$, matrix H_a being determined as in the isotropic case. Metric $\mathcal{M}(K)$ can be defined as the intersection of the three metrics $\mathcal{M}(x)$ for $x = a, b$, and c (cf. Frey and George, 2000).

5.3.3 A surface-based approach

In the previous sections, we proposed a majoration about the interpolation error on the basis of the Hessian of the solution that has been directly used to obtain the size constraints about the mesh elements. In the present section, we propose a new approach by considering an appropriate Cartesian surface.

Let Ω be the computational domain, let $\mathcal{T}(\Omega)$ be a mesh of Ω , and let $u(\Omega)$ be the physical solution obtained in Ω via mesh $\mathcal{T}(\Omega)$. The pair $(\mathcal{T}(\Omega), u(\Omega))$ allows us to define a Cartesian surface $\Sigma_u(\mathcal{T})$ (we assume u to be a scalar function). Given $\Sigma_u(\mathcal{T})$, the problem of minimizing the interpolation error consists in defining an (optimal) mesh $\mathcal{T}_{\text{opt}}(\Omega)$ in Ω such that surface $\Sigma_u(\mathcal{T}_{\text{opt}})$ is as smooth as possible. To this end, we propose a local characterization of the surface near a vertex. Two methods are introduced: the first using the local deformation allows for an isotropic adaptation, while the other, using the local curvature, results in an anisotropic adaptation.

Local deformation of a surface

The basic idea consists in a local characterization of the deviation (at order 0) of surface mesh $\Sigma_u(\mathcal{T})$ near a vertex with respect to a reference plane, in specific, the tangent plane of the surface at this vertex. This deviation can be evaluated by considering the Hessian along the normal to the surface (e.g. the second fundamental form).

Let P be a vertex in the solution surface $\Sigma_u(\mathcal{T})$. Locally, near P , this surface has a parameterized form $\sigma(x, y)$, (x, y) being the parameters, with $P = \sigma(0, 0)$. Using a Taylor expansion at order 2 of σ near P , results in

$$\begin{aligned} \sigma(x, y) &= \sigma(0, 0) + \sigma'_x x + \sigma'_y y \\ &\quad + \frac{1}{2}(\sigma''_{xx} x^2 + 2\sigma''_{xy} xy + \sigma''_{yy} y^2) + o(x^2 + y^2) \end{aligned}$$

where $e = (1, 1, 1)$. If $\nu(P)$ stands for the normal to the surface at P , then quantity $\langle \nu(P), (\sigma(x, y) - \sigma(0, 0)) \rangle$ represents the gap from point $\sigma(x, y)$ to the tangent plane in P and can be written as

$$\begin{aligned} &\frac{1}{2}(\langle \nu(P), \sigma''_{xx} \rangle x^2 + 2\langle \nu(P), \sigma''_{xy} \rangle xy \\ &\quad + \langle \nu(P), \sigma''_{yy} \rangle y^2) + o(x^2 + y^2) \end{aligned}$$

which is therefore proportional to the second fundamental form of the surface when $x^2 + y^2$ is small enough.

The local deformation of the surface at P is defined as the maximal gap of the vertices adjacent to P to the tangent plane of the surface at P . If (P_i) denotes those vertices, then the local deformation $\varepsilon(P)$ of the surface at P is given by

$$\varepsilon(P) = \max_i \langle \nu(P), \overrightarrow{PP_i} \rangle$$

Therefore, the optimal mesh of $\Sigma_u(\mathcal{T})$ for Ω is a mesh where the size at all nodes p is inversely proportional to $\varepsilon(P)$ where $P = (p, u(p))$. Formally speaking, the optimal size $h_{\text{opt}}(p)$ associated with node p is written as

$$h_{\text{opt}} = h(p) \frac{\varepsilon}{\varepsilon(P)}$$

where ε is the given threshold, and $h(p)$ is the size of the elements near p in mesh $\mathcal{T}(\Omega)$.

As can be seen, the local deformation is a rather easy way to characterize the local deviation of the surface, which does not involve computing the explicit computation of the Hessian of the solution. The only drawback in this measure is that it allows only an isotropic adaptation. In the same context (minimizing the local deviation), using the curvature allows an analysis of this deviation, which is both more precise and anisotropic.

Local curvature of a surface

Analyzing the local curvature of the surface related to the solution also makes it possible to minimize the deviation (order 1) from the tangent planes of the solution that interpolate the exact solution. Indeed, while considering the construction of isotropic surface meshes, we have shown in Borouchaki *et al.* (2001) how these two deviations, of order 0 and 1, are bounded by a given threshold and how the element size at all vertices is proportional to the minimal radius of curvature. Let $P = (p, u(p))$ be a vertex in $\Sigma_u(\mathcal{T})$, let $\rho_1(P)$ and $\rho_2(P)$ with $\rho_1(P) \leq \rho_2(P)$ be the two principal radii of curvature, and let $(\vec{e}_1^*(P), \vec{e}_2^*(P))$ be the corresponding unit principal directions. The ideal size at P is

$$h_{\text{opt}}^*(P) = \gamma \rho_1(P)$$

where γ is a factor related to the specified threshold about the deviation. This size is defined in the tangent plane at the surface at P . Let us consider the frame $(P, \vec{e}_1^*(P), \vec{e}_2^*(P))$ in this plane, if $h_{\text{opt}}^*(P)$ reads $h_{\text{opt}}^*(P) = h_1^* \vec{e}_1^*(P) + h_2^* \vec{e}_2^*(P)$ in this frame then the constraint in size at P can be written in $(P, \vec{e}_1^*(P), \vec{e}_2^*(P))$ by

$$\left(h_1^* \quad h_2^* \right) \frac{\mathcal{I}_2}{\sqrt{2} \rho_1^*(P)} \begin{pmatrix} h_1^* \\ h_2^* \end{pmatrix} = 1$$

which is the equation of a circle centered at P in the tangent plane at the surface at P . By means of an orthogonal projection of this circle in the plane of Ω , we obtain the size constraint at p . If $\vec{v}_1^*(p)$ and $\vec{v}_2^*(p)$ are the orthogonal projections of $\vec{e}_1^*(P)$ and $\vec{e}_2^*(P)$ in the plane of Ω , then this size constraint in frame $(p, \vec{v}_1^*, \vec{v}_2^*)$ ($\vec{v}_1^* = (1, 0)$ et $\vec{v}_2^* = (0, 1)$) is given by

$$\begin{aligned} &\left(h_1 \quad h_2 \right)^t \left(\left(\vec{v}_1^*(p) \quad \vec{v}_2^*(p) \right)^{-1} \right)^t \frac{\mathcal{I}_2}{\sqrt{2} \rho_1^*(P)} \\ &\quad \times \left(\vec{v}_1^*(p) \quad \vec{v}_2^*(p) \right)^{-1} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = 1 \end{aligned}$$

where (h_1, h_2) are the coordinates in the frame $(p, \vec{v}_1^*, \vec{v}_2^*)$ of the projection of $h_{\text{opt}}^*(P)$ in the plane of Ω . This relationship defines a metric (which is in general anisotropic) at p .

The metric previously defined may lead to a large number of elements due to the isotropic nature of the elements in the surface. In order to minimize this number of elements, and for anisotropic meshing purposes, a similar relationship involving the two principal radii of curvature can be exhibited (Frey and George, 2000). In such a case, the ideal size of the surface element is given using a so-called geometric metric, which, at vertex P of $\Sigma_u(\mathcal{T})$, is written as

$$\begin{aligned} &\left(h_1^* \quad h_2^* \right) \begin{pmatrix} 1 & 0 \\ \gamma^2 \rho_1^*(P) & 1 \end{pmatrix} \\ &\quad \times \begin{pmatrix} h_1^* \\ h_2^* \end{pmatrix} = 1 \end{aligned}$$

where γ is a factor related to the given threshold about the deviation and $\eta(\gamma, \rho_1(P), \rho_2(P))$ is a function related to γ , $\rho_1(P)$, and $\rho_2(P)$, which ensures a similar deviation along the two principal directions. This relationship generally describes an ellipse in the tangent plane of the surface at P , which includes the circle obtained in the isotropic case. Similarly, the corresponding metric at p is obtained after a projection of this ellipse in the plane of Ω .

In practice, computing the local curvature at all vertices in this surface first leads to computing the normal (then the gradient) by means of a weighted average of the unit normal at the adjacent elements. Then, in the local frame (the tangent plane together with the normal), we construct a quadratic centered at this vertex, which passes at best through the adjacent vertices, after which we consider locally the Hessian to be that of this quadric. Finally, using the gradient and the Hessian at the nodes of $\mathcal{T}(\Omega)$, we compute the principal curvatures and directions at all vertices in the surface $\Sigma_n(\mathcal{T})$.

5.4 Solution interpolation

In the proposed adaptive schema, it is necessary to interpolate the current solution from the former mesh to the new one in order to continue the computation at a given stage.

This step involves interpolating solution $S_i(\Omega)$ associated with mesh $\mathcal{T}_i(\Omega)$ on $\mathcal{T}_{i+1}(\Omega)$. In the case where there are no physical constraints in the PDE problem under solution, the interpolation problem is written as a simple optimization problem like

$$\min \|S_{i+1}(\Omega) - S_i(\Omega)\|$$

where $\|\cdot\|$ is a norm, for instance, a L_2 or a H_1 Sobolev norm, and each solution S is associated with its corresponding mesh. The solution of this minimization problem necessitates computing the intersection of mesh $\mathcal{T}_i(\Omega)$ and mesh $\mathcal{T}_{i+1}(\Omega)$. In cases where the physical constraints must be considered, the underlying problem is a constrained minimization problem. This is well understood for linear physical constrained operators (Bank, 1997).

However, in practice, a simple linear interpolation of $S_i(\Omega)$ on $\mathcal{T}_{i+1}(\Omega)$ allows a solution that is close to the ideal targeted solution. This linear interpolation does not require complex computations such as explicitly computing the mesh intersections.

6 LARGE-SIZE PROBLEM, PARALLELISM AND ADAPTIVITY

Large-size problems are mainly encountered in some complex CFD calculations even where adaptive methods are used. Parallel technologies are therefore the only solution to handle such problems. Parallelism may be included in various computational steps. Mesh construction together with solution methods may benefit to some extent from parallelism. The first aspect involves in constructing the mesh in parallel, while the second point concerns parallelizing solvers (see Chapter 20, Chapter 22 of this Volume).

6.1 Parallel meshing methods

There are essentially two different ways to construct a mesh in parallel following an a posteriori or an a priori approach. In the first approach, a given mesh of the domain is subdivided into a number of meshed regions such that the number of elements is well balanced from one processor to the other while minimizing the number of nodes at the processor (region) interfaces. Other criteria related to the problem in hand can be added to the above classical requirements. Various strategies have been proposed to achieve these goals. After this domain decomposition, first the interfaces between regions are meshed in parallel, then regions are considered for being meshed in parallel (Shostko and Löhner, 1995; Aliabadi and Tezduyar, 1995; Weatherill *et al.*, 1998; Löhner, 2001). Such an approach involves a serial-meshing method (while parallel meshing methods also exist). The second approach no longer considers a domain mesh but, in its place, defines the subdivision using only the boundary discretizations. In this way, defining an interface requires finding a surface that passes through a contour part of the domain boundary decomposition (Galtier, 1997). Then this surface is meshed, thus completing the definition of the regions, after which each region is meshed in parallel.

The main difficulty of the first approach is the need for an initial mesh of the full domain, while, in the second approach, the key-point is the proper definition of surface interfaces. Conversely, the first approach allows for a simple definition of the interfaces (as part of the volume mesh faces) and the second approach avoids meshing the full domain (with a coarse mesh in general).

6.2 Parallel solvers

While various paradigms are used in the solvers, the main issue is the proper management of the interfaces from region to region. This requires communication between meshes. In this respect, minimizing the node interface as well as insuring some degree of smoothness at the interface level is of great importance. Also of interest is load balancing to avoid having idle processors (i.e. waiting for others). Balancing the load necessitates, a priori, evaluating the number of elements in each region, which is quite easy when using the first approach, whereas it has proved to be a tedious task in the second approach.



Figure 3. Uniform and geometric meshes of a part of a 747 boeing model.

7 MESHING FOR MOVING BOUNDARY PROBLEMS

Moving boundary problems mainly occur in solid mechanical engineering specifically in stamping and forming processes. To some extent, simulating the trajectory of a body in some CFD problems or modeling the behavior of a semiconductor device in etching processes lead to closely related problems. In solid mechanical problems, the geometry of the mechanical part is not known during the process, and it is subjected to large deformations. In such a case, the remeshing step is a crucial part of the simulation. For CFD concerns, the geometry is known in advance, but rigid movements apply, and, in this sense, such a problem can be seen as a classical remeshing problem. Meshing for moving boundary problems where large deformations are applied to the geometry of the domain can be addressed using two classes of methods. The first involves mesh optimization that considers a topologically correct (in terms of connectivity) mesh that is geometrically invalid (negative volume or overlapping elements). The other class makes use of a full two-step remeshing procedure; remeshing the domain boundary prior to remeshing the domain itself. The following briefly discusses the latter method, while the reader is referred to Coupez (1991) for the first approach.

7.1 Moving problems in two dimensions

Let Ω be the mechanical part defined from its boundary Γ assumed to be made up of piecewise linear line segments $\mathcal{T}(\Gamma)$. This boundary discretization can be obtained from a CAD definition of Γ . However, this definition is no longer useful when defining the geometry at a further step of the deformation process (the final deformation is assumed to be the summation of small step changes). The remeshing can be applied after each increment of deformation as follows:



Figure 4. Uniform mesh of a car seat (a) and its flattening map (b); data courtesy of LECTRA. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>.

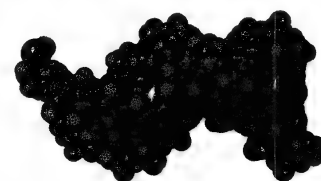


Figure 5. Regular mesh of a DNA molecule. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>.

- Definition of the new geometry $\mathcal{G}(\Gamma')$ after deformation
- Geometric error estimation (deviation of the current discretization $\mathcal{T}(\Gamma')$ from the new geometry $\mathcal{G}(\Gamma')$) resulting in a size map $\mathcal{H}_G(\Gamma')$ used to govern the discretization of Γ'

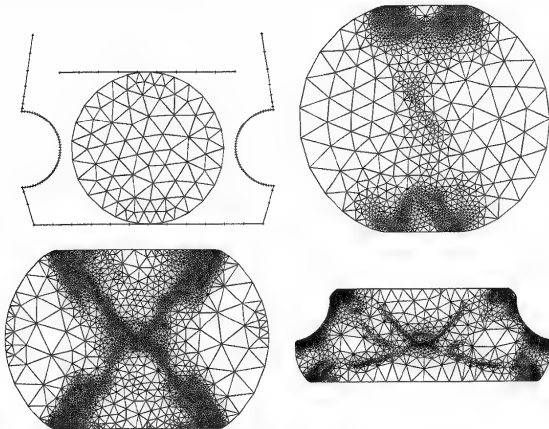


Figure 6. Forming of an asymmetrical mechanical part (top left, initial configuration).

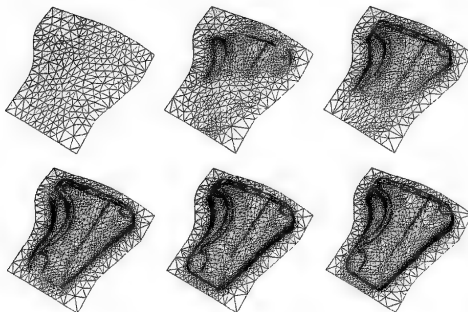


Figure 7. Stamping of a sheet of metal, from the planar sheet to the stamped result.

- Physical error estimation (deviation of the current solution $S(\Omega)$ from an ideal solution assumed to be smooth enough) that results in a size map $\mathcal{H}_\phi(\Omega)$ serving to govern the remeshing of Ω .
- Definition of the full size map $\mathcal{H}(\Omega)$ by merging $\mathcal{H}_\phi(\Gamma)$ and $\mathcal{H}_\phi(\Omega)$.
- Adaptive discretization of Γ with respect to $\mathcal{H}(\Omega)$.
- Adaptive remeshing of Ω with respect to $\mathcal{H}(\Omega)$.

The deformations of the geometry include *free* together with *bounded* deformations. The first type is a free deformation due to a mechanical constraint (for instance, equilibrium conditions). In this case, the new geometry of the part after deformation is only defined by the new positions of the boundary nodes together with their connections. The second type is a deformation limited by a contact with a second domain whose geometry is fixed (the tool is assumed to be rigid). In this case, the part takes the geometric shape of the tool, and thus its geometry after deformation is that of the tool.

Geometric error estimation is based on the evaluation of the curvature of the boundary at each node (for a free node, this curvature is that of the current boundary, while for a bounded node, the curvature is that of the related part of the tool in contact). Physical error estimation is based on the interpolation error and can be accessed by computing the discrete Hessian of the current solution.

Merging two size maps involves constructing a unique map where the size is the minimum of the sizes in the two given maps.

With this material, the remeshing procedures follow the same aspect as in a classical adaptive remeshing scheme (see above).

7.2 Moving problems in three dimensions

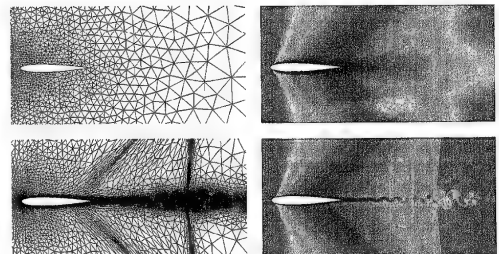
In this case, the boundary is made up of linear triangular elements and the remeshing scheme proposed in the previous section applies. However, some points are much more tedious, including

- identifying the bounded nodes,
- computing the discrete Hessian part of the physical error estimate,
- the full three-dimensional nature of the adaptive remeshing process.

For the sake of simplicity, the bounded nodes can be dealt with like the free nodes.

8 APPLICATION EXAMPLES

A number of application examples are given to demonstrate the approaches we have proposed. Figure 3 displays a uniform mesh and a geometric mesh constructed on a 747 boeing model defined by a uniform fine grid made up of quads. After using this grid to define a geometry (by means of Coons patches), we meet a parametric surface-meshing problem. Figure 4 demonstrates the construction of a regular mesh for a series of CAD patches. Figure 5 shows the mesh of a DNA molecule (by means of a Connolly surface) where the geometry is defined by the constituent atoms (e.g. a series of intersecting spheres). Figures 6 and 7 give examples about forming and stamping problems (e.g. moving boundary problems). Figure 8 illustrates two stages of a mesh for a transonic calculation in two dimensions,

Figure 8. Transonic flow around a Naca0012 wing, (initial and adapted) meshes and isodensities. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

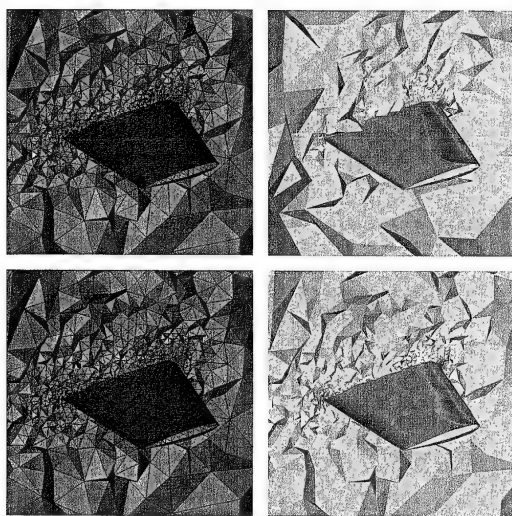


Figure 9. Transonic flow around a wing in three dimensions, cut through the tet mesh and isodensities (initial and adapted meshes). A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

while Figure 9 considers an example in three dimensions. Figure 10 shows an example in biomedical engineering. Then, Figure 11 gives an example of mesh simplification, which is a useful method for various engineering problems as well as for image compression or compact data storage.

9 CONCLUSIONS

In this chapter, we have discussed mesh-generation methods and mesh-adaptivity issues for automated planar, surface, and volume meshing. After a review of the classical mesh-generation methods, we have considered adaptive schemes where the solution is to be accurately captured. To this end, meshing techniques have been revisited to be capable of completing high-quality meshes conforming to these features. Error estimates have therefore been introduced to

analyze the solution field at a given stage prior to being used to complete adapted meshes. Some details about large-size meshes and moving boundary problems have been given. Application examples have been shown to demonstrate the various approaches proposed throughout the chapter.

While mature in a number of engineering fields, current meshing technologies still need to be investigated to handle nonsimplicial elements (quads, hexes (still a challenge to date), ...) as well as nonlinear elements (quadratic or higher degrees). Surprisingly, surface meshing has not been particularly well addressed so far. Robust implementation for anisotropic meshing in three dimensions is still a field of intensive work. Meshing problems that include colliding regions are certainly a pertinent subject of future investigations. Meshes with billions of elements also lead to interesting topics on massively parallel strategies.

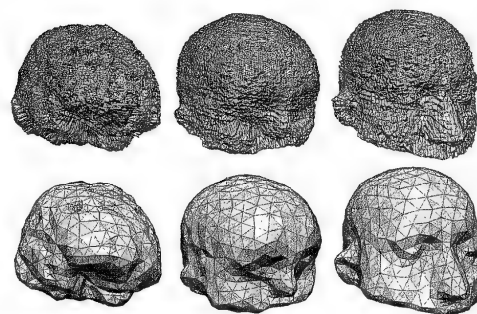


Figure 10. Initial dense meshes of a brain, a cranial bone and a scalp and corresponding simplified meshes (Hausdorff distance). A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>



Figure 11. Simplified meshes of Lucy statue and corresponding enlargements. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

REFERENCES

- Aliabadi S and Tezduyar T. Parallel fluid dynamics computations in aerospace applications. *Int. J. Numer. Methods Fluids*. 1995; 21:783–805.
- Allwright SE. Techniques in multiblock domain decomposition and surface grid generation. In *Grid Generation in Computational Fluid Mechanics*, Sengupta S, Thompson JP, Eiseleman PR and Hauser J (eds). Pineridge Press: Swansea, 1988; 559–568.
- Anglada MV, Garcia NP and Crosa PB. Directional adaptive surface triangulation. *Comput. Aided Geometric Des.* 1999; 16:107–126.
- D'Azevedo EF and Simpson B. On optimal triangular meshes for minimizing the gradient error. *Numer. Math.* 1991; 59(4):321–348.
- Babuska I and Aziz A. On the angle condition in the finite element method. *SIAM J. Numer. Anal.* 1976; 13:214–227.
- Bank RE. Mesh smoothing using a posteriori estimates. *SIAM J. Numer. Anal.* 1997; 34(3):979–997.
- Berzins M. Mesh quality: a function of geometry, error estimates or both? *Eng. Comput.* 1999; 15:236–247.
- Boissonnat JD and Yvinec M. *Algorithmic Geometry*. Cambridge University Press, 1997.
- Borouchaki H, Chapelle D, George PL, Laug et P and Frey P. Estimateur d'erreur géométrique et adaptation. *Mailage et*

- adaptation, *Traité Mécanique et Ingénierie des Matériaux*, Hermès-Lavoisier, Paris, in French, 2001; 279–310.
- Borouchaki H, George PL, Hecht F, Laug P and Salte E. Delaunay mesh generation governed by metric specifications. Part I. Algorithms. *Finite Elem. Anal. Des.* 1997; 25:61–83.
- Borouchaki H, Hecht F and Frey PJ. Mesh gradation control. *Int. J. Numer. Methods Eng.* 1997; 43:1143–1165.
- Carey GF. *Computational Grids: Generation, Adaptation and Solution Strategies*. Taylor & Francis, 1997.
- Ciarlet PG. Basic error estimates for elliptic problems. In *Handbook of Numerical Analysis*, Vol. II, Ciarlet PG and Lions JL (eds). North Holland, 1991; 17–352.
- Cook WA. Body oriented coordinates for generating 3-dimensional meshes. *Int. J. Numer. Methods Eng.* 1974; 8:27–43.
- Coupez T. Grandes transformations et remaillage automatique. Thèse ENSMP, CEMEF, 1991.
- Fortin M. Estimation a posteriori et adaptation de maillages. *Revue européenne des éléments finis* 2000; 9(4).
- Frey PJ and Borouchaki H. Geometric surface mesh optimization. *Comput. Visual. Sci.* 1998; 1:113–121.
- Frey PJ and George PL. *Mesh Generation*. Hermès: Oxford, also in french, 2000.
- Galtier J. Structures de données irrégulières et architectures haute performance. Une étude du calcul numérique intensif par le partitionnement de grappes. Thèse, University of Versailles, 1997.
- George A. Computer Implementation of the Finite Element Method. PhD thesis, Dept. of Computer Science, Stanford Univ., 1971.
- George PL. *Automatic Mesh Generation. Applications to Finite Element Methods*. Wiley, 1991.
- George PL and Hermeline F. Delaunay's mesh of a convex polyhedron in dimension d. Application to arbitrary polyhedra. *Int. J. Numer. Methods Eng.* 1992; 33:975–995.
- George PL (eds). *Maillage et adaptation*. *Traité Mécanique et Ingénierie des Matériaux*, Hermès-Lavoisier, in french. Paris, 2001.
- George PL and Borouchaki H. *Delaunay Triangulation and Meshing, Application to Finite Element*. Hermès: Paris, also in french, 1998.
- Hermeline F. Une méthode automatique de maillage en dimension n. Thèse Paris VI, Université Paris VI, Paris, 1980.
- Hughes TJR. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Prentice-Hall: Englewood Cliffs, 1998.
- Joe B. Construction of three-dimensional Delaunay triangulations using local transformations. *Comput. Aided Geom. Des.* 1991; 8:123–142.
- Kaupp P and Steinberg S. *The Fundamentals of Grid Generation*. CRC press, 1993.
- Lo SH. A new mesh generation scheme for arbitrary planar domains. *Int. J. Numer. Methods Eng.* 1985; 21:1403–1426.
- Lo SH. Automatic mesh generation and adaptation by using contours. *Int. J. Numer. Methods Eng.* 1991; 31:689–707.
- Löhner R and Parikh P. Three-dimensional grid generation by the advancing front method. *Int. J. Numer. Methods Fluids* 1988; 8:1135–1149.
- Löhner R. Automatic unstructured grid generators. *Finite Elem. Anal. Des.* 1997; 25(3–4):111–134.
- Löhner R. A parallel advancing front grid generation scheme. *Int. J. Numer. Methods Eng.* 2001; 51:663–678.
- Marcum DL and Weatherill NP. Unstructured grid generation using iterative point insertion and local reconnection. *AIAA J.* 1995; 33(9):1619–1625.
- Peraire J, Vahdati M, Morgan K and Zienkiewicz OC. Adaptive remeshing for compressible flow computations. *J. Comput. Phys.* 1987; 72:449–466.
- Peraire J, Peiro J and Morgan K. Adaptive remeshing for three-dimensional compressible flow computations. *J. Comput. Phys.* 1992; 103:269–285.
- Preparata FP and Shamos MI. *Computational Geometry, an Introduction*. Springer-Verlag, 1985.
- Raviart et PA and Thomas JM. (1988), *Introduction à l'analyse numérique des équations aux dérivées partielles*. Masson: Paris.
- Rippa S. Long and thin triangles can be good for linear interpolation. *SIAM J. Numer. Anal.* 1992; 29:257–270.
- Sheng X and Hirsch BE. Triangulation of trimmed surfaces in parametric space. *Comput. Aided Des.* 1992; 24(8):437–444.
- Shoemaker A and Löhner R. Three-dimensional parallel unstructured grid generation. *Int. J. Numer. Methods Eng.* 1995; 38:905–925.
- Thompson JF, Soni BK and Weatherill NP. *Handbook of Grid Generation*. CRC Press, 1999.
- Thompson JF, Warsi ZUA and Mastin CW. *Numerical Grids Generation, Foundations and Applications*. North Holland, 1985.
- Vallet MG. Génération de maillages éléments finis anisotropes et adaptatifs. Thèse Université ParisVI, 1992.
- Van Phai N. Automatic mesh generation with tetrahedron elements. *Int. J. Numer. Methods Eng.* 1982; 18:237–289.
- Verfürth R. *A Review of a Posteriori Error Estimation and Adaptive Refinement Techniques*. Wiley Teubner, 1996.
- Watson DF. Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes. *Comput. J.* 1981; 24:167–172.
- Weatherill NP and Hassan O. Efficient three-dimensional Delaunay triangulation with automatic point creation and imposed boundary constraints. *Int. J. Numer. Methods Eng.* 1994; 37:2005–2039.
- Weatherill NP, Said R and Morgan K. The construction of large unstructured grids by parallel Delaunay grid generation. In *Proc. 6th Int. Conf. on Numerical Grid Generation in Computational Field Simulation*, Cross M, Eiseman P, Hauser J, Soni BK and Thompson JF (eds). M.S.U., 1998; 53–78.
- Yerry MA and Shephard MS. A modified quadtree approach to finite element mesh generation. *IEEE Comput. Graph. Appl.* 1983; 3(1):39–46.
- Yerry MA and Shephard MS. Automatic three-dimensional mesh generation by the modified-octree technique. *Int. J. Numer. Methods Eng.* 1984; 20:1965–1990.
- Zienkiewicz OC. *The Finite Element Method*. McGraw-Hill: London, 1977.
- Borouchaki H and George PL. Quality mesh generation. *C.R. Acad. Sci. Paris* Concise review paper, L328, Serie II-b 2000; 505–518.
- George PL. Automatic mesh generation and finite element computation. In *Handbook of Numerical Analysis, Vol IV, Finite Element methods (Part 2), Numerical Methods for Solids (Part 2)*, Ciarlet PG and Lions JL (eds). North Holland, 1996; 69–190.

FURTHER READING

Apel T. *Anisotropic Finite Element: Local Estimates and Applications*. Wiley Teubner, 1999.

Chapter 18

Computational Visualization

William J. Schroeder¹ and Mark S. Shephard²

¹Kitware, Inc., Clifton Park, NY, USA

²Rensselaer Polytechnic Institute, Troy, NY, USA

| | |
|--|-----|
| 1 Introduction | 525 |
| 2 Data Forms | 528 |
| 3 Visualization Algorithms | 531 |
| 4 Volume Rendering | 541 |
| 5 Methods in Large Data Visualization | 542 |
| 6 Taxonomy for Data Visualization Systems | 543 |
| 7 Interfacing the Computational System with the Visualization System | 546 |
| References | 548 |

1 INTRODUCTION

This section defines visualization and introduces a taxonomy of different types of visualization. Uses of visualization in the computational sciences are also described.

1.1 What is visualization?

In its broadest definition, visualization is the transformation of data or information into sensory input perceivable by the human observer (Schroeder, Martin and Lorensen, 2003). The purpose of visualization is to engage the human perceptual system in such a way as to transmit pertinent information to the analyst as efficiently as possible. Unlike automated processes such as artificial intelligence that attempt

to produce results independent of the human observer, the visualization process presumes the observer's existence and engages him/her directly in the exploration of the problem. For this reason, the most effective visualization techniques are interactive, that is, they respond to commands from the observer in real time with minimal latency and data rates of approximately 5 Hz or greater. Batch or off-line production of visualization results is also important, but these are typically preprogrammed once the analyst knows where to look and what to look for.

By definition, visualization techniques encompass sensory input of vision, sound, touch (haptics), taste, and smell. Visual representations remain far and away the most widely used techniques, while sound- and haptic-based methods have been used in some applications with limited success. Visualization techniques based on taste and olfactory input remain experimental.

1.2 Terminology

Different terminologies are used to describe visualization. *Scientific visualization* is the formal name given to the field in computer science that encompasses user interface, data representation and processing algorithms, visual representations, and other sensory presentation such as sound or touch (McCormick, DeFanti and Brown, 1987). Scientific visualization is generally used in the context of spatial-temporal domains (such as those found in computational mechanics). The term *data visualization* is another phrase used to describe visualization. Data visualization is generally interpreted to be more general than scientific visualization, since it implies treatment of data sources beyond the sciences and engineering. Such data sources include financial,

marketing, or business data. In addition, the term 'data visualization' is broad enough to include application of statistical methods and other standard data analysis techniques (Rosenbium *et al.*, 1994). Another recently emerging term is *information visualization*. This field endeavors to visualize abstract information such as hypertext documents on the World Wide Web, directory/file structures on a computer, or abstract data structures (The First Information Visualization Symposium, 1995). A major challenge facing information visualization researchers is to develop coordinate systems, transformation methods, or structures that meaningfully organize and represent data.

Another way to classify visualization technology is to examine the context in which the data exists. If the data is spatial-temporal in nature (up to three spatial coordinates and the time dimension), then typically methods from scientific visualization are used. If the data exists in higher-dimensional spaces, or abstract spaces, then methods from information visualization are used. This distinction is important, because the human perceptual system is highly tuned to space-time relationships. Data expressed in this coordinate system is inherently understood with little need for explanation. Visualization of abstract data typically requires extensive explanations as to what is being viewed and what the display paradigm is. This is not to say that there is no overlap between scientific and information visualization – often the first step in the information visualization process is to project abstract data into the spatial-temporal domain, and then use the methods of scientific visualization to view the results. The projection process can be quite complex, involving methods of statistical graphics, data mining, and other techniques, or it may be as simple as selecting a lower-dimensional subset of the original data.

The term 'visualization' may be used to mean different operations depending on the particular context. *Geometry visualization* is the viewing of geometric models, meshes or grids, or other information representing the topology and geometry of the computational domain. Geometry visualization often includes modeling operations such as cutting, clipping, or extracting portions of the domain, and may include supplemental operations such as collision detection and measurement. *Results visualization* typically combines the viewing of geometry in conjunction with attribute data, for example, coloring the surface of an aircraft wing with pressure values. Attribute data is typically classified as scalar (single-valued), vector (n -component vector), or tensor (general matrix of values). Techniques in scientific visualization also include a modeling component, for example, producing stream surfaces from a three-dimensional vector field. In this example, the vector

data is used to control the generation of a surface following the rules of fluid flow (a surface tangent to the flow at all points).

1.3 The role of visualization in computational methods

Visualization techniques can be used in all phases of the computational process: preprocessing (defining input geometry, boundary conditions, and loads); solution (numerical solution); and postprocessing (viewing, interacting, and analyzing results). The following sections provide an overview of the tasks involved in each phase.

1.3.1 Preprocessing

Modern computational systems rely heavily on visualization and graphics techniques to assist the analyst to define the input geometry, specify boundary conditions, apply loads to the model and/or generate the computational grid. Geometry definition requires modeling the domain, including using abstract representations (e.g. 2-manifolds embedded in 3-space to represent shell-like structures). This is typically an interactive activity requiring graphical input to define, shape, and edit the geometric representation. Loading and boundary conditions are often applied using additional geometric primitives to indicate the region, direction, and magnitude of application. Mesh generation is often an automatic process once the geometry, loading, and boundary conditions have been defined. However, interactive input may be required to indicate regions of higher mesh density, or to perform validation of the mesh subsequent to generation. In some cases, manual mesh generation is still performed. This may require interactive methods to decompose the domain into topologically regular blocks, specify gradation density in each block, and control mesh layout.

1.3.2 Postprocessing

Visualization is best known for its use in postprocessing the results of analysis. Geometry visualization is used to inspect the computational grid or underlying geometric representation if the problem produces evolving geometry (e.g. shape optimization). Results visualization is used to view scalar, vector, and tensor fields, often in the context of the geometry (spatial domain) of the problem. Other important techniques include probing data to produce 1-D or 2-D plots; clipping, cutting, and extracting data to focus on particular regions of the domain; animating results over time or design iterations to view the evolution of the solution; and comparing results across different analyses.

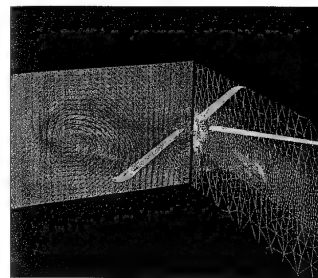


Figure 1. Visualization of computational analysis. Here, flow over a rotating blade is shown. (Courtesy SCORE/Rensselaer.) A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

Many of these techniques will be described in more detail later in this chapter.

1.3.3 Analytical steering

Visualization techniques are employed during the solution phase to monitor solution convergence or view intermediate results. Some systems provide methods to *analytically steer* the solution process. For example, while monitoring the numerical solution process, it is possible to adjust time steps or the computational grid to accelerate convergence or improve accuracy. In practice, analytical steering may be implemented by producing intermediate solution files, or directly linking the postprocessing module to the solver. When generating intermediate files, standard postprocessing tools and techniques are used to visualize the results at which point the analyst may adjust input parameters or restart the solution with modified input. Relatively few systems provide run-time linking directly to the solution solver. While this is arguably the better approach, it requires rewriting software to integrate the visualization system into the numerical solver.

1.4 3-D computer graphics

Underlying most visualization techniques are methods based on computer graphics. Because of recent efforts to accelerate 3-D rendering on relatively inexpensive graphics hardware – motivated principally by gaming, video, and

other forms of entertainment – graphics display methods tend to follow particular patterns based on the capabilities of the hardware (i.e. accelerated polygon rendering with textures). This in turn drives the implementation of visualization algorithms. For example, most algorithms currently produce linear primitives such as points, lines, and polygons. Representing geometry with these graphics primitives will generally produce responsive, interactive visualizations. In addition, modern graphics hardware supports features such as texture mapping. Texture mapping can be used to great benefit for coloring surfaces based on data value, 'cutting' data using textures with transparency, and performing efficient volume rendering with 3-D textures. As a result, many areas of visualization research languish (e.g. methods based on higher-order, nonlinear primitives) in preference to techniques that produce higher rates of interaction such as polygon and texture-based rendering techniques.

In addition, because methods in computer graphics are generally used, the analyst must be aware of the inherent limitations and terminology used in the rendering process, lighting models, and camera models. In the next few sections, we present some basic information. For more information, please see books on computer graphics such as Foley *et al.* (1990) and Watt (1993).

1.4.1 Rendering

In 3-D computer graphics, a single image is produced by rendering a scene (Figure 1). The scene consists of geometry with associated attribute data (color or color index, surface normals, texture coordinates, etc.); a camera to project the geometry onto the view plane (two cameras are used for stereo rendering); and lights to illuminate the geometry. In addition, transformation matrices (normally 4×4 homogeneous matrices) are used to position the geometry, lights, and camera. Various properties are used to control the effect of lighting and the appearance of the geometry (described in the following section).

While a wide variety of rendering techniques are known in computer graphics, in visualization two basic approaches are used: surface-based and volume rendering. Surface rendering projects linear geometric primitives such as triangles or polygons onto the view plane. Primitives are rasterized into pixel values of depth and color using a scan conversion process (linear interpolation from the polygon edges into the interior). Surface properties controlling the color of the surface and the lighting model affect the appearance of the geometry. Texture mapping is often used to project an image onto the surface geometry, with the placement of the texture map controlled by texture coordinates. Surface transparency can be specified to control the visibility

of the interior of objects and its relationship to other surfaces. Volume rendering – covered in more detail later in this chapter – takes into account the interior of objects to see into or through them, producing X-ray-like images showing interior structure and data variation.

A major concern of the rendering process is to properly sort primitives in the direction of the view vector. The result is an image with correct depth occlusion and/or properly blended color if the object is translucent. Typically, this is performed in hardware using the z-buffer and color buffer to retain the depth and color of the pixel closest (in depth, or z-direction) to the view position. However, proper blending of translucent objects, or use of volume rendering, requires an ordering of the primitives.

1.4.2 Lighting model

The lighting model generally considers three types of lighting effects: ambient, diffuse, and specular lighting. Ambient lighting is an overall illumination independent of surface orientation and relationship to a particular light. Diffuse lighting, or Lambertian reflection, takes into account the orientation of the surface relative to the incident light vector. Specular reflection represents the direct reflection of light from an object's surface toward the observer. Specular reflection takes into account both the incident light vector, the view vector and the surface normal, whereas diffuse lighting takes into account only the relation of the light vector to the surface normal.

Surface normals play an important role in the lighting model. Flat shading, where the normal across a primitive (i.e. polygon) is taken as a constant, results in a faceted appearance. Gouraud shading computes color at each vertex and then uses linear interpolation during scan conversion to color the interior of the polygon. Phong shading interpolates the normal from the vertices of the polygon to produce a smoother light variation. Figure 2 shows the effect of flat shading (constant normal per polygonal cell) versus Gouraud shading, which interpolates the normal across each polygon.

1.4.3 Camera model

While a complete treatment of the projection of geometry from 3-D into 2-D is beyond the scope of this chapter, there are a few key concepts of which the visualization practitioner must be aware. Projection methods are typically of two types: perspective and orthographic projection. Perspective projection takes into account the view angle of the observer to produce perspective effects (e.g. closer objects appear bigger, parallel lines converge at infinity). Orthographic projection does not include perspective effects, so

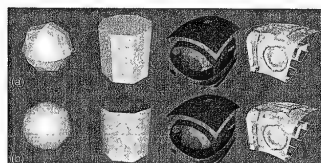


Figure 2. Gouraud (b) versus flat shading (a). (Courtesy of Kitware, Inc. Taken from the book *The Visualization Toolkit: An Object-Oriented Approach to 3-D Graphics Third Edition* ISBN-1-930934-07-6.) A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

a rendered object remains the same size no matter how close or far from the camera it is (a view scale is used to zoom in and out).

In computer graphics, the camera model defines a local coordinate system consisting of the camera position, the view plane normal (which may be different than the direction of projection resulting in projection shearing effects), the view up vector, and near and far clipping planes. The clipping planes are oriented perpendicular to the view direction. All objects before the near plane, and beyond the far plane are culled prior to rendering, which can be used to great benefit to focus on a narrow slice of data.

Stereo viewing is achieved using two cameras – one for the left and one for the right eye. The angle of separation between the eyes controls the binocular effect. However, too large an angle results in eyestrain and difficulty in fusing the two images.

2 DATA FORMS

A variety of data forms are used in the computational sciences. Visualizing these data requires matching the computational data to the data forms found in visualization.

2.1 Overview

There are two distinct data forms that compose a visualization dataset. The first is the spatial representation, or geometric and topological structure of the data. In the computational science, the structure is typically represented by the computational grid or mesh. The second is the data attributes – sometimes referred to as data fields – associated with the structure. The dataset structure is composed of cells and points (sometimes referred to as the elements

and nodes). Cells define a topological type and define the relationship of the cells points. Points define a geometric coordinate and fix the cells in space. Most visualization algorithms assume that the field data is associated with either the points and/or cells of the dataset. Very few visualization techniques treat data associated with intermediate topological levels such as the edges or faces of a 3-D cell (e.g. tetrahedron). Typically, the data structures used to represent the topology and geometry of a visualization dataset are more compact and of limited functionality as compared to the computational system. This is primarily due to the emphasis that visualization places on interactive display and successful management of large data.

2.2 Spatial representations

The dataset structure is classified according to whether the structure is regular or irregular, and similarly, whether it is implicitly or explicitly represented. These classifications apply to both the geometry and topology of the dataset. A dataset with regular geometry has point coordinates x_i that are specified by a function $\vec{x}_i = f(i)$. A dataset with regular topology is one where the number and type of cells in the dataset, and the neighbors to each cell are known implicitly from a specification of the dataset. Practically speaking, the difference between regular and irregular datasets is that regular topology and geometry can be implicitly defined whereas irregular data requires an explicit representation of geometry and/or topology. These considerations are particularly important in implementation because implicit representations require far fewer computational resources (both in memory and CPU) than explicit representations. On the other hand, irregular data tends to have greater flexibility in its ability to adaptively represent the domain, and in some cases, may require fewer computational resources than regular data of equivalent resolution.

2.2.1 Regular data

The most common type of regular data (regular both in topology and geometry) is images (2-D) and volumes (3-D) – regular lattices of points arranged parallel to the coordinate axes. Image processing and volume rendering are just two of many disciplines devoted to the processing and display of this type of data. The principal advantage to this form is its simple representation (origin, inter-pixel/voxel spacing, and lattice dimensions). This manifests itself as compact algorithms that are easily parallelized as compared to irregular data forms.

Structured grids are another form of regular data. Such grids are regular in topology, but irregular in geometry.

That is, the dataset is described by sample dimensions in each coordinate axis (e.g. $10 \times 20 \times 30$ grid in 3-D) along with a vector of point coordinates explicitly specifying the position of each point in the grid. Such grids are often collected together to form multiblock datasets, where each block of data corresponds to a different grid.

Rectilinear grids are regular in topology, and like image/volume data, axis-aligned. However, the point coordinates are semiregular. A single vector of coordinate values per coordinate axis is required.

2.2.2 Irregular data

Unstructured grids represent the most general form of irregular data. Both the points and cells must be represented explicitly. Tetrahedral meshes, or meshes consisting of mixed element types are examples of unstructured grids. Points are typically represented via a vector of coordinate values; cells are represented by a type specification plus a connectivity array. Because visualization systems generally represent data as compactly as possible, the intermediate topological and geometric hierarchy is often not represented explicitly; rather it is derived from the point/cell relationships. In some cases, the intermediate hierarchy is represented on an as-needed basis.

A subset of the general unstructured grid is often referred to as graphics data, or polygonal datasets. These datasets are composed of linear graphics primitives such as vertices, lines, polygons, and triangle strips. They are an important form because graphics systems are typically optimized for display of such primitives. Also, many viewing operations of even 3-D data require only the surface of the dataset since interior cells are hidden. Hence, polygonal data often serves as an intermediate form between the analysts' data and the graphical subsystem.

Another common form of irregular data is unorganized point sets. For example, laser digitizers can scan millions of points from the surface of an object in a few seconds to produce such data. It is also common to subsample a general n -dimensional dataset into a set of three-dimensional points, and then use typical visualization techniques to explore the subset.

2.2.3 Other representations

Computational scientists frequently employ adaptive meshes such as quadrees, octrees, and AMR grids to represent the problem domain. Often, these are transformed in the visualization system to similar representations tuned for minimal memory requirements or interactive processing. For example, the branch-on-need-octree (BONO) (Wilhelms and Van Gelder, 1992) and

interval tree (Livnat, Shen and Johnson, 1996) represent single-valued scalar data to accelerate the generation of isocontours. Other forms, such as AMR grids (Berger and Olinger, 1984), may be triangulated (on the fly or as a preprocessing step) to create unstructured grids.

Higher-order meshes such as p -basis finite elements are underrepresented in visualization systems. Many systems support quadratic – and some even cubic – isoparametric basis functions on simplices and rectangular elements. However, most visualization systems require pretriangulation of higher-order basis, or may automatically triangulate higher-order cells into linear primitives that can be processed by conventional visualization algorithms. It is important that the analysts understand whether such data conversions take place, and if so, what the implications are on the accuracy, memory requirements, and computational resources of the visualization process. Ideally, such considerations should take place initially as part of the overall strategy for solving a particular problem. (See Section 1.7 for more information about interfacing meshes to the visualization system.)

2.3 Dataset attributes

Dataset attributes are associated with the structure of the dataset, typically with the points and cells. Visualization researchers typically classify visualization techniques according to the type of attribute data they operate on, as well as the underlying structure of the dataset. The general categories of algorithm research are scalar fields, vector fields, and tensor fields. A whole host of additional algorithms use combinations of these techniques, or use methods from computer graphics and computational geometry to create particular types of visualizations. For example, glyphing is a data-driven modeling operation as described in the following section.

2.3.1 Scalar fields

A set of single-valued data values is referred to as a scalar field, or simply scalars. Temperature, pressure, density, a component of displacement or stress, factor of safety, and so on are all examples of scalar values. There is a close correspondence between scalars and colors – colors being a vector tuple of grayscale, RGB (red-green-blue), or RGBA (red-green-blue-alpha with alpha a measure of transparency). The relationship between scalars and colors is via a lookup table (scalar value indexes into a table of colors) or a function known as a transfer function $c_i = f(s_i)$ that maps a scalar value s_i into a unique color c_i . This relationship forms the basis of many visualization techniques

including the color mapping and volume rendering methods described shortly.

2.3.2 Vector fields

An n -tuple of values representing a direction and magnitude defines a vector, where n is typically (2, 3) in two- and three-dimensional space. Velocity, momentum, displacement, direction, and gradients form typical vector fields.

2.3.3 Tensor fields

Tensors are complex mathematical generalizations of vectors and matrices. A tensor of rank k can be considered a k -dimensional table. A tensor of rank 0 is a scalar, rank 1 is a vector, and rank 3 is a three-dimensional rectangular array. Existing visualization algorithms focus on rank 2 tensors, that is, 3×3 matrices such as strain and stress tensors.

2.3.4 Graphics attributes

Because visualization is inherently tied to computer graphics, attributes related to the rendering of data are also important. This includes surface normals and texture coordinates. Surface normals (a normalized direction vector) are used in the shading process to show the effects of lighting as we saw previously. Texture mapping is used to apply detail to the surface of objects (the texture is generated from the data) or to model objects (transparent or translucent textures can be used to cut away or window into data).

2.3.5 Time

Time is generally used to organize the previously described attributes into separate time steps. Visualizations are performed for each step, and a sequence of steps are arranged to form animations. Some algorithms – such as streakline generation and time domain-based image processing algorithms – may use time to produce specialized visualizations.

2.3.6 General fields

A general collection of data may be referred to as a field. For example, the collection of stresses, strains, and displacements at nodal points form a field of results from a materials analysis. Other information, such as run identifier, optimization values, and material properties can be thrown into the mix. The point is that many visualization algorithms can represent general fields. However, in order to visualize the data, the field is winnowed down into scalars, vectors, tensors, or one of the other attributes described previously and then visualized. In some cases, fields may be combined

to form irregular point sets and visualized using techniques appropriate to that class of datasets.

2.4 Cells and data interpolation

As previously indicated, the current abstraction for visualization systems assumes that the dataset is composed of cells. These cells are typically linear (lines, triangles, polygons, tetrahedra) or products of linear functions (quadrilateral, hexahedral). The primary purpose of the cell is to provide an interpolating function within the spatial extent that the cell encompasses. That is, the cell provides a continuous field of values from the discrete set of points at which the data field is computed or sampled. Typically, the interpolating functions are the standard isoparametric functions found in finite element analysis. The cell also serves to partition the domain into topologically regular subdomains over which geometric computation and searching can be accelerated.

Current visualization systems provide limited support for cell types. Computational scientists using advanced higher-order p -type finite element formulations, or those using adaptive meshes with complex types (e.g. octants arbitrarily subdivided in an octree) will certainly encounter these limitations. The standard recourse in such cases is to subdivide the cells into the appropriate linear primitives supported by the visualization system. Of course, this subdivision must be performed carefully to avoid losing information. Note, however, that excessive subdivision may result in excessive numbers of primitives resulting in slow interaction rates. (Section 1.7 contains more information about interfacing meshes to the visualization system.)

The cell data abstraction is an outcome of several factors. Graphics systems are optimized for linear interpolation of data values across simple primitives such as polygons; ultimately, all data must be transformed into this form if interactive visualization is to be used. Second, data tends to be discrete – that is, data values are known at fixed positions – while rendering techniques are designed for continuous surfaces. It is particularly important to understand the effects of interpolation, and to be aware of the differences in interpolation function found between the computational system and the visualization system.

3 VISUALIZATION ALGORITHMS

Visualization algorithms are classified according to the attribute and dataset type they treat. The dataset type generally impacts the speed and complexity of the algorithms, although some algorithms are structure-specific. For example, three-dimensional regular lattice data (e.g. volumes)

can be readily subsampled into planes and lines of data. Such techniques do not exist for irregular data.

3.1 Scalar fields

The most used visualization algorithms are those for scalar fields.

3.1.1 Color mapping

Color mapping transforms scalar values into colors, and applies the colors to a surface geometry that is then rendered (Figure 3). The colors may be shades of gray (grayscale), variation across a particular color hue, an arbitrary color function or any of these combined with alpha transparency values. The geometry may range from spatial extractions of cells and points to surfaces created by extracting the boundary of the domain or a particular subset of cells. The surface may be arbitrary such as a probe surface or cut plane.

Color mapping is often implemented using a lookup table. Given a particular scalar data range (s_{\min} , s_{\max}) and

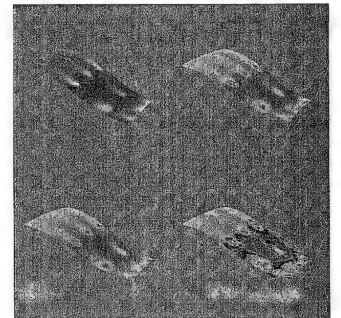


Figure 3. Color mapping can produce dramatically different results depending on the choice of lookup table (or transfer function). Visualization must necessarily consider the human perceptual system. As a result, the computational scientist must carefully consider the effect of perceptual factors on the target audience. (Courtesy of Kivware, Inc. Taken from the book *The Visualization Toolkit: An Object-Oriented Approach to 3-D Graphics Third Edition* ISBN-1-930934-07-6.) A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

a scalar value s_i , a linear lookup table computes an index into the table using the function

$$\text{index} = \frac{s_i - s_{\min}}{s_{\max} - s_{\min}}$$

where s_i is assumed to lie between (s_{\min}, s_{\max}) , and if not, is clamped to these values. The table color values are defined by the user. In many systems, predefined tables or simple functions are used to define the color table.

More complex mappings are also available. *Transfer functions* are continuous versions of lookup tables that map a scalar value into a color. Separate transfer functions may be available for each component of an RGB color and alpha transparency. Volume rendering (see Section 1.4) typically uses transfer functions in implementation.

3.1.2 Isocontours

One of the most common visualization techniques generates isocontours from scalar fields. An isocontour is the line (2-D) or surface (3-D) defined by all points taking on a particular scalar value (the contour value C)

$$f(x, y, z) = C \quad (1)$$

While there are a multitude of methods for computing isocontours, in general, they are based on edge interpolation. An edge intersects an isosurface when the data values associated with its two end points are simultaneously above and below a specified contour value. Linear interpolation is then used to produce the isosurface intersection point. Various methods are then used to connect the intersection points into edges, faces, and ultimately, $(n-1)$ -dimensional cells (assuming that the dataset is n -dimensional).

A widely used isocontouring technique is the so-called *marching cubes* algorithm (Lorensen and Cline, 1987). Marching cubes is simple to implement, fast, and lends itself to parallel implementation. Furthermore, while the original algorithm was described on voxels (the cells of a regular volume dataset), it is easily extended to other cells of varying topology, for example, triangles, quadrilaterals, and tetrahedra. The essence of the algorithm is that given a particular isocontour value C , the data values on the points forming the cell produce a finite number of combinations with respect to C . That is, each point may be classified in one of two states: (1) its data value may be greater than or equal to C , or (2) its data value may be less than C . For a hexahedron, 256 (2^8) finite combinations exist. As a result, it is possible to create a *case table* that explicitly represents the topology of the resulting triangulation of the cell; that is, the number of points, edges, and triangles that approximate the isosurface. The algorithm proceeds by determining

the case of each cell, indexing into the case table, and producing the prescribed intersection points and triangles. Figure 4 shows a reduced case table that combines the various cases – related by symmetry or rotation – into 15 distinct configurations. Note, however, that in practice the full 256-element table is used to insure that the resulting isosurface is manifold (i.e. contains no holes). The case table includes arbitrary decisions regarding the triangulation of hexahedral faces with ambiguous configurations (see Figure 5). The full case table insures that the decisions are made consistently. Methods are available to better estimate what cases to use (Nielsen and Hammam, 1991); but ultimately the data is sampled at discrete points and the ambiguity cannot be rigorously resolved.

Duplicate points generated along the same edge from neighboring voxels are typically combined using a point merging operation. Normals may also be computed from interpolation of point gradient values. The point gradient values in turn are computed using a central difference scheme easily computed from the 3-D lattice structure of the volume.

Because marching cubes is case-table-based, it is both fast and robust. However, the algorithm must visit all voxels, most of which do not contain the isosurface. A variety of acceleration techniques such as span space (Livnat, Shen and Johnson, 1996) can be used to insure that only those cells containing isosurface are visited. The interval tree keeps track of cells according to min/max scalar values. If a particular isocontour value falls within a particular min/max range of a cell, then that cell must be processed. Empty cells are ignored.

Marching cubes is easily extended to other cell types, including irregular cell types such as tetrahedra. The primary difference is the manner in which point normals are computed. In irregular data forms, the computation of point normals is usually deferred until after the isosurface is created. Normals can then be computed from the surface directly.

3.1.3 Carpet plots

Displacement of a surface as a function of scalar value is known as a *carpet plot*. For example, topographic maps displaced in the z -direction as a function of elevation create two and a half dimensional terrain maps. A similar approach is used to create carpet plots of pressure across an airplane wing, for example. The magnitude of displacement is further controlled using a scale factor to prevent excessive or imperceptible displacement. The same technique can be used for an arbitrary surface. Note that the displacement direction vector – which is typically taken as the surface normal – varies rapidly on highly curved surfaces. The

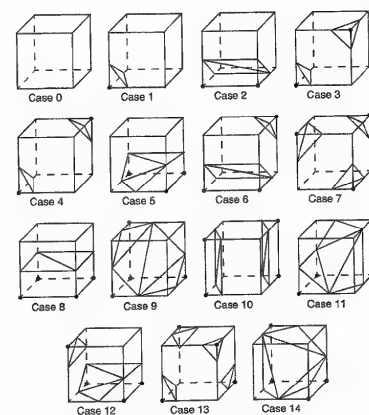


Figure 4. Marching cubes case table. In practice, all 256 possible cases are used to insure that the surface remains manifold.

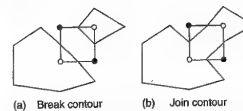


Figure 5. Choosing a particular contour case will break (a) or join (b) the current contour. The case table must be designed to insure that this decision is made consistently across all cases.

resulting carpet plot can be confusing in such situations. The plot surface may even self-intersect in some cases.

3.2 Vector fields

Vector fields are also widely found in the computational sciences. Vector displays show vector magnitude or direction, or both.

3.2.1 Vector glyphs

Probably, the most common vector displays are those based on glyphs. Typically, a glyph (such as a small

arrow) is oriented in the direction of the local vector, and may be colored or scaled to indicate magnitude of the vector. Glyphed vector fields are also referred to as *hedgehogs* because of their resemblance to the small spiked animals. Care must be used when scaling glyphs since uniform scaling of 3-D objects yields $O(n^2)$ change in surface area and $O(n^3)$ change in volume. These effects may mislead the consumer of the visualization.

3.2.2 Displacement maps

Another approach converts vectors to scalar values, and then uses methods of scalar visualization to view the result. For example, surface vectors can be converted to scalars by computing the dot product between the vector and the local surface normal. Color mapping can then be used to shade the surface to produce a *displacement map*. Figure 6 shows the result of this technique applied to a beam in vibration. The visualization clearly indicates the regions of positive and negative displacement, and the nodal lines (lines of zero displacement) between the regions.



Figure 6. A displacement map of a beam in vibration. (Courtesy of Kitware, Inc. Taken from the book *The Visualization Toolkit An Object-Oriented Approach to 3-D Graphics Third Edition* ISBN-1-930934-07-6.) A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

3.2.3 Vector displacement plots

Similar to scalar carpet plots, *vector displacement plots* are used to illustrate vector displacement. The original dataset geometry is deformed using the vector field, usually in combination with a scale factor to control the amount of distortion. These plots are particularly effective when rapidly animated by varying the scale factor using a sinusoidal or saw tooth function.

3.2.4 Particle advection

The structure of the vector field can be explored using interactive techniques such as *particle advection*. This is an extension of the hedgehog technique, where a point is moved a small distance over a time period dt . In other words, if velocity $\vec{V} = dx/dt$, then the displacement of a point is

$$dx = \vec{V} dt \quad (2)$$



Figure 7. Time animation of a point C . Although the spacing between points varies, the time increment between each point is constant.

This suggests an extension to our previous techniques: repeatedly displace points over many time steps. Figure 7 shows such an approach. Beginning with a sphere S centered about some point C , S is repeatedly moved to generate the bubbles shown. The eye tends to trace out a path by connecting the bubbles, giving the observer a qualitative understanding of the fluid flow in that area. The bubbles may be displayed as an animation over time (giving the illusion of motion) or as a multiple exposure sequence (giving the appearance of a path). This is referred to as *particle advection*.

Such an approach can be misused. For one thing, the velocity at a point is instantaneous. Once we move away from the point, the velocity is likely to change. Equation (2) above assumes that the velocity is constant over the entire step. By taking large steps, we are likely to jump over changes in the velocity. Using smaller steps, we will end in a different position. Thus, the choice of step size is a critical parameter in constructing accurate visualization of particle paths in a vector field.

To evaluate equation (2), we can express it as an integral:

$$\vec{x}(t) = \int_t \vec{V} dt \quad (3)$$

Although this form cannot be solved analytically for most real world data, its solution can be approximated using numerical integration techniques. Accurate numerical integration is a topic beyond the scope of this book, but it is known that the accuracy of the integration is a function of the step size dt . The simplest form of numerical integration is Euler's method,

$$\vec{x}_{i+1} = \vec{x}_i + \vec{V}_i \Delta t \quad (4)$$

where the position at time \vec{x}_{i+1} is the vector sum of the previous position plus the instantaneous velocity times the incremental time step Δt .

Euler's method has error on the order of $O(\Delta t^2)$, which is not accurate enough for many applications. One such example is shown in Figure 8. The velocity field describes perfect rotation about a central point. Using Euler's method, we find that we will always diverge, and instead of generating circles, will generate spirals instead.

A better approach is the Runge-Kutta technique of second order (Conte and de Boor, 1972). This is given by

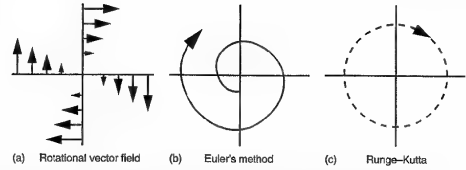


Figure 8. Euler's integration (b) and Runge-Kutta integration of order 2 (c) applied to uniform rotational vector field (a). Euler's method will always diverge.

the expression

$$\vec{x}_{i+1} = \vec{x}_i + \frac{\Delta t}{2} (\vec{V}_i + \vec{V}_{i+1}) \quad (5)$$

where the velocity \vec{V}_{i+1} is computed using Euler's method. The error of this method is $O(\Delta t^3)$. Compared to Euler's method, the Runge-Kutta technique enables larger integration step at the expense of one additional function evaluation.

3.2.5 Numerical integration

Integration formulas require repeated transformation from world coordinates to local cell coordinates and as a result are computationally demanding. The local vector field is computed via cell interpolation (in local coordinates) whereas the spatial displacement is computed in global coordinates. There are two important steps we can take to improve performance.

1. *Improving search procedures.* There are two distinct types of searches. Initially, the starting location of the particle (i.e. to find what cell contains the point) must be established using a global search procedure. Once the initial location of the point is determined, an incremental search procedure can be used. Incremental searching is efficient because the motion of the point is limited within a single cell, or at most across a cell boundary. Thus, the search space is greatly reduced, and the incremental search is faster relative to the global search.
2. *Coordinate transformation.* The cost of a coordinate transformation from global to local coordinates can be reduced if either of the following conditions is true: the local and global coordinate systems are identical with one another (or vary by a simply rigid body transform), or if the vector field can be transformed from global space to local coordinate space. The image data coordinate system is an example of local coordinates that

are parallel to global coordinates, hence global to local coordinate transformation can be greatly accelerated. If the vector field is transformed into local coordinates (either as a preprocessing step or on a cell-by-cell basis), then the integration can proceed completely in local space. Once the integration path is computed, selected points along the path can be transformed into global space for the sake of visualization.

3.2.6 Particle traces, streamlines, and streaklines

A natural extension to the methods of the previous section is to connect the point position $\vec{x}(t)$ over many time steps. The result is a numerical approximation to a particle trace represented as a line. Depending on whether the flow is steady or time varying, we can define three related line representation schemes for vector fields.

- *Particle traces* are trajectories traced by fluid particles over time.
- *Streaklines* are the set of particle traces at a particular time t_i that have previously passed through a specified point x_i .
- *Streamlines* are integral curves along a curve s satisfying the equation

$$s = \int_t \vec{V} ds, \quad \text{with } s = s(x, t) \quad (6)$$

for a particular time t .

Streamlines, streaklines, and particle traces are equivalent to one another if the flow is steady. In time-varying flow, a given streamline exists only at one moment in time. Visualization systems generally provide facilities to compute particle traces. However, if time is fixed, the same facility can be used to compute streamlines. In general, this visualization algorithm is referred to as streamline generation, but it is important to understand the differences when the vector field is time varying.

3.2.7 Advanced methods

The vector field integration techniques described previously lend themselves to a variety of related methods. A simple extension to the streamline is to wrap the line with a tube. The tube radius may vary according to mass flow (Schroeder, Volpe and Lorensen, 1991). That is, assuming incompressible flow with no shear, the radius of the tube can vary according to the scalar function vector magnitude. Then the equation

$$r(\vec{v}) = r_{\max} \sqrt{\frac{|\vec{v}|}{|\vec{v}_{\min}|}} \quad (7)$$

relates an area of constant mass flow, where the radius of the tube at any point $r(\vec{v})$ is a function of the maximum radius r_{\max} and minimum velocity along the tube \vec{v}_{\min} .

Another common streamline technique widens the line to create a ribbon or surface. One method to create a stream-surface generates adjacent streamlines and then bridges the lines with a ruled surface. This technique works well as long as the streamlines remain relatively close to one another. If separation occurs, so that the streamlines diverge, the resulting surface will not accurately represent the flow, because we expect the surface to be everywhere tangential to the vector field (i.e. definition of streamline). The ruled surface connecting two widely separated streamlines does not generally satisfy this requirement; thus the surface must adaptively adjust to local flow conditions.

A streamribbon can also be calculated by attaching a ribbon to the streamline and rotating it with the local streamwise vorticity. Vorticity $\vec{\omega}$ is the measure of rotation of the vector field, expressed as a vector quantity: a direction (axis of rotation) and magnitude (amount of rotation). Streamwise vorticity Ω is the projection of $\vec{\omega}$ along the instantaneous velocity vector, \vec{v} . Said in another way, streamwise vorticity is the rotation of the vector field around the streamline defined as follows.

$$\Omega = \frac{\vec{v} \cdot \vec{\omega}}{|\vec{v}| |\vec{\omega}|} \quad (8)$$

The amount of twisting of the streamribbon approximates the streamwise vorticity.

A streamsurface is a collection of an infinite number of streamlines passing through a *base curve*. The base curve, or *rake*, defines the starting points for the streamlines. If the base curve is closed (e.g. a circle), the surface is closed and a streamtube results. Thus, streamribbons are specialized types of streamsurfaces with a narrow width compared to length.

Compared to vector icons or streamlines, streamsurfaces provide additional information about the structure of the

vector field. Any point on the streamsurface is tangent to the velocity vector. Consequently, taking an example from fluid flow, no fluid can pass through the surface. Streamtubes are then representations of constant mass flux. Streamsurfaces show vector field structure better than streamlines or vector glyphs because they do not require visual interpolation across icons.

Streamsurfaces can be computed by generating a set of streamlines from a user-specified rake. A polygonal mesh is then constructed by connecting adjacent streamlines. One difficulty with this approach is that local vector field divergence can cause streamlines to separate. Separation can introduce large errors into the surface, or possibly cause self-intersection, which is not physically possible.

3.2.8 Vector field topology

Vector fields have a complex structure characterized by special features called *critical points* (Globus, Levit and Lasinski, 1991; Helman and Hesselink, 1991). Critical points are locations in the vector field where the local vector magnitude goes to zero and the vector direction becomes undefined. At these points, the vector field either converges or diverges, and/or local circulation around the point occurs (Figure 9).

A number of visualization techniques have been developed to construct vector field topology from an analysis of critical points. These techniques provide a global understanding of the field, including points of *attachment* and *detachment* and field *vortices*. Using a fluid flow analogy, points of attachment and detachment occur on the surface of an object where the tangential component of the vector field goes to zero, and the flow is perpendicular to the surface. Thus, streamlines will begin or end at these points. There is no common definition for a vortex, but generally speaking, vortices are regions of relatively concentrated vorticity (e.g. flow rotation). The study of vortices is important because they represent areas of energy loss, or can have significant impact on downstream flow conditions (e.g. trailing vortices behind large aircraft).

One useful visualization technique creates vector field skeletons that divide the vector field into separate regions. Within each region, the vector field is topologically equivalent to uniform flow. These skeletons are created by locating critical points, and then connecting the critical points with streamlines. In 3-D vector field analysis, this technique can be applied to the surface of objects to locate lines of flow separation and attachment and other important flow features. Also, in general 3-D flow, the regions of uniform flow are separated by surfaces, and creation of 3-D flow skeletons is a current research topic.

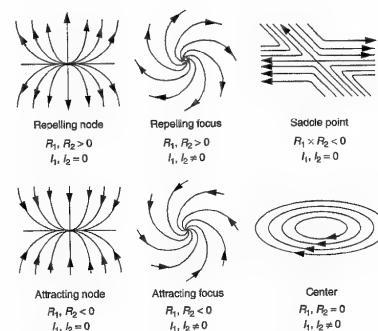


Figure 9. Critical points in two dimensions. The real part of the eigenvalues (R_1, R_2) of the matrix of first derivatives control the attraction or repulsion of the vector field. The imaginary part of the eigenvalues (I_1, I_2) controls the rotation.

3.3 Tensor fields

Tensor visualization is an active area of research. There are few techniques for tensor visualization other than 3×3 real symmetric tensors. Such tensors are used to describe the state of strain or stress in a 3-D material. The well-known stress and strain tensors for an elastic material are shown in Figure 10.

Recall that a 3×3 real symmetric matrix can be characterized by the three eigenvectors and three eigenvalues of the matrix. The eigenvectors form a 3-D coordinate system whose axes are mutually perpendicular. In some applications, particularly the study of materials, these axes also are referred to as the principal axes of the tensor and are

$$\begin{aligned} \text{(a) Stress tensor} & \quad \begin{bmatrix} \sigma_x & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \sigma_y & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \sigma_z \end{bmatrix} \\ \text{(b) Strain tensor} & \quad \begin{bmatrix} \frac{\partial u}{\partial x} & \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) & \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right) \\ \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) & \frac{\partial v}{\partial y} & \left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}\right) \\ \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right) & \left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}\right) & \frac{\partial w}{\partial z} \end{bmatrix} \end{aligned}$$

Figure 10. Stress and strain tensors. Normal stresses in the x - y - z coordinate directions indicated as $\sigma_x, \sigma_y, \sigma_z$, shear stresses indicated as τ_{ij} . Material displacement represented by u, v, w components.

physically significant (e.g. directions of normal stress and no shear stress). Eigenvalues are physically significant as well. In the study of vibration, eigenvalues correspond to the resonant frequencies of a structure, and the eigenvectors are the associated mode shapes.

We can express the eigenvectors of the 3×3 system as

$$\vec{v}_i = \lambda_i \vec{e}_i, \quad \text{with } i = 1, 2, 3 \quad (9)$$

with \vec{e}_i , a unit vector in the direction of the eigenvalue, and λ_i , the eigenvalues of the system. If we order eigenvalues such that

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \quad (10)$$

then we refer to the corresponding eigenvectors \vec{v}_1, \vec{v}_2 , and \vec{v}_3 as the *major*, *medium*, and *minor* eigenvectors.

3.3.1 Tensor ellipsoids

This leads us to the tensor ellipsoid technique for the visualization of real, symmetric 3×3 matrices. The first step is to extract eigenvalues and eigenvectors as described in the previous section. Since eigenvectors are known to be orthogonal, the eigenvectors form a local coordinate system. These axes can be taken as the *minor*, *medium*, and *major* axes of an ellipsoid. Thus, the shape and orientation of the ellipsoid represent the relative size of the eigenvalues and the orientation of the eigenvectors. In Figure 11 we

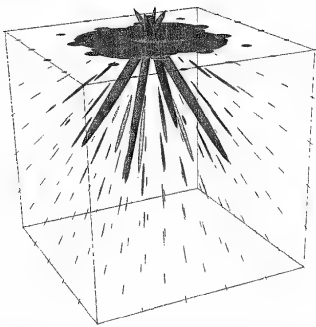


Figure 11. Tensor ellipsoids. (Courtesy of Kitware, Inc. Taken from the book *The Visualization Toolkit An Object-Oriented Approach to 3-D Graphics Third Edition* ISBN-1-930934-07-6.) A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

visualize the analytical results of Boussinesq's problem from Saada. Note that tensor ellipsoids and tensor axes are a form of *glyph* specialized to tensor visualization.

3.3.2 Hyperstreamlines

Hyperstreamlines are constructed by creating a streamline through one of the three eigenfields, and then sweeping a geometric primitive along the streamline (Delfmarcelle and Hesselink, 1993). Typically, an ellipse is used as the geometric primitive, where the remaining two eigenvectors define the major and minor axes of the ellipse (Figure 12). Sweeping the ellipse along the eigenfield streamline results

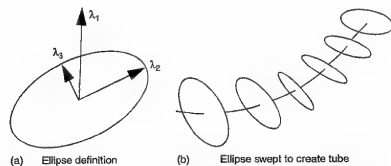


Figure 12. Creation of hyperstreamlines. An ellipse is swept along a streamline of the eigenfield. Major/minor axes of the ellipse are controlled by the other two eigenvectors.

in a tubular shape. Another useful generating geometric primitive is a cross. The length and orientation of the arms of the cross are controlled by two of the eigenvectors. Sweeping the cross results in a helical shape since the eigenvectors (and therefore the arms of the cross) will typically rotate in the tensor field. Figure 13 shows an example of hyperstreamlines. The data is from a point load applied to a semiinfinite domain. Compare this figure to Figure 11 that used tensor ellipsoids to visualize the same data. Notice that there is less clutter and more information available from the hyperstreamline visualization.

3.4 Extraction and modelling

Visualization inherently involves modeling operations and the extraction of subregions of data. For example, an analyst may wish to extract a portion of the data whose scalar values lie within a particular scalar range (i.e. threshold the data). Or, operations that clip and cut may provide a view into the interior of a complex 3-D dataset. Extraction and clipping is generally performed to limit the data to a region of interest, either to gain interactive performance from large data sets, or to eliminate visual distraction from less important regions of the data.

Data extraction and modeling operations often transform the structure of the data on which they operate. For example, if all (voxel) cells from a regular volumetric dataset and contained within an ellipsoidal region are extracted, the resulting structure is not regular, and therefore, not a volume. Alternatively, extracting a region of interest (ROI) from a volume will result in another regular volume. Operations that modify the structure of data may have a dramatic impact on the performance and computational requirements of a visualization process.

3.4.1 Geometric extraction and implicit functions

Geometric extraction produces cells and/or points that lie within a domain – typically a spatial domain. The test for

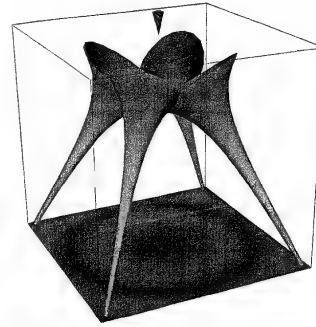


Figure 13. Example of hyperstreamlines. The four hyperstreamlines shown are integrated along the minor principal stress axis. A plane (colored with a different lookup table) is also shown. (Courtesy of Kitware, Inc. Taken from the book *The Visualization Toolkit An Object-Oriented Approach to 3-D Graphics Third Edition* ISBN-1-930934-07-6.) A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

inclusion may include cells that are completely within the domain, or cells with one or more points lying within the domain (i.e. partial inclusion). The domain can be defined in a number of ways, including using bounding boxes or combinations of clipping planes. In particular, the use of implicit functions – including boolean combinations of implicit functions – is a powerful, simple way to define complex domains. An implicit function has the form

$$F(\vec{x}) = 0 \quad (11)$$

where $F(\vec{x}) < 0$ is *inside* and $F(\vec{x}) > 0$ is *outside* the implicit function. The family of implicit functions includes planes, spheres, cones, ellipsoids, and a variety of other simple shapes. Using boolean combinations – union, intersection, and difference – it is possible to create complex domain definitions using these simple shapes.

Implicit functions have the property that they convert a position \vec{x} into a scalar value s via equation (11). Thus, any scalar technique described previously (e.g. isocontouring) can be used in combination with implicit functions. This forms the basis of several techniques such as thresholding, cutting, and clipping described in the following sections.

3.4.2 Thresholding

Thresholding is a technique to extract subregions of data based on attribute values. For example, we may wish to extract all cells whose scalar values fall within a specified range. Generally, thresholding is performed using scalar values because any attribute type can be converted into a single value using the appropriate evaluation function (e.g. vector dot product, etc.). In the previous section, we saw how implicit functions can be used to convert a position \vec{x} into a scalar value.

There are several ways in which to implement the thresholding operation. One is to geometrically extract cells satisfying the threshold into a new, output dataset. This has the benefit of reducing data size but may convert a regular dataset into an irregular one as described previously. Another approach is to create a 'blanking' array to indicate which cells and/or points are visible. During the rendering process, only the visible entities are actually drawn. The benefit to this approach is that the regularity of the data is preserved at the expense of the additional memory. Finally, another interesting approach derived from computer graphics is to use transparent textures to eliminate invisible data. That is, a special texture map consisting of a transparent region and an opaque region is used in combination with texture coordinates computed from the thresholding operation. The benefit of the texture approach is that the structure of the dataset is not modified and the thresholding function can be changed rapidly by modifying the relatively small texture map. Furthermore, since modern graphics hardware supports texture maps, it is possible to produce interactive thresholding on large datasets with this approach.

3.4.3 Topological extraction

Topological extraction selects cells and points based on topological criterion such as id or adjacency to a particular feature such as an edge or face. Dataset boundaries of manifold 3-D datasets can be determined by extracting all faces that are used by one cell (in the sense that a face is used when it forms the boundary of a 3-D cell). Often, topological and geometric operations are combined – sharp 'feature' edges (determined by a geometric operation comparing surface normals on faces using an edge) and the faces connected to these edges (a topological adjacency operation) can be displayed to examine the qualities of a computational mesh. Other common operations include selecting a point and displaying all edges, faces, and/or cells using that point.

3.4.4 Probing and spatial location

Probing is used to transform structure, reduce data size, reduce topological dimension, and focus on particular features of a dataset. The fundamental probing operation determines the data values (e.g. attribute values) at a particular point in a dataset. Hence, this sampling operation involves two distinct steps: locating the point within the structure of the dataset (e.g. determining which cell contains the probe point) and interpolating data values from the containing cell. This operation inherently reduces the topological dimension of a dataset down to that of a zero-dimensional point. However, by arranging a set of points in a particular structure, it is possible to produce new datasets of varying dimension and structure. Furthermore, if the number of points in the probe is less than that of the original dataset, data reduction is performed.

Typical examples of probing include determining data values at a point of interest (i.e. producing a numerical value), producing x - y plots along a line or curve, producing surface plots, color maps, or images from a regular array of points, and converting irregular unstructured data into regular volumetric data (for the purposes of volume rendering). Like any sampling function, probing must be used carefully to avoid undersampling and oversampling data. Undersampling may produce incorrect visualizations that miss important data features; oversampling may produce the illusion of accuracy and require excessive computational resources.

3.4.5 Cutting

The cutting operation reduces an n -dimensional dataset into an $(n-1)$ -dimensional surface (the *cutting surface*). For example, a 3-D dataset may be cut by a plane to produce a 2-D surface on which the original data attributes have been interpolated.

Cutting operations can be conveniently implemented using isocontouring methods in conjunction with implicit functions. The implicit function is used to describe the cutting surface. During the cutting operation, each point of each cell is evaluated through the implicit function to produce values that are above, below, and equal to zero. An isosurface of value $F(\vec{x}) = 0$ is then extracted to produce the cutting surface. Isocontouring and cutting are equivalent in the sense that both are controlled by scalar values. The difference is that the cutting operation produces scalar values by evaluating point locations through an implicit function.

One of the advantages of the cutting operation as compared to probing is that the resolution of the cut surface is directly related to the resolution of the underlying data through which the cut surface passes.

3.4.6 Clipping

The clipping operation produces an n -dimensional dataset from an $(n-1)$ -dimensional *clipping surface* applied to an n -dimensional input dataset. The structure of the data is often transformed by this operation. For example, clipping a 3-D regular volume with a plane produces an irregular output that will consist of tetrahedra. Similar to cutting operations, clipping surfaces are often described by implicit functions. However, data values (i.e. scalar values) can also be used to define the clipping surface.

Clipping is relatively easy to produce in cells of dimension two or less. Case tables similar to that of the marching cubes isocontouring algorithm are designed to produce linear cell types, such as lines, triangles, and quadrilaterals. In dimensions three and higher, it is difficult to design case tables that produce consistent meshes with cell types closed under the set tetrahedron, hexahedron, pyramid, and wedge. For example, while tetrahedra can be clipped to produce new tetrahedra and wedges, hexahedra require careful pre-tetrahedrization to produce compatible meshes. (Consistent meshes are those that satisfy the compatibility conditions described by Luebke *et al.* (2003) – no T-junctions or free faces.)

3.4.7 Glyphing

Glyphing is a general-purpose visualization technique that can be used in endless variety. We have already seen how glyphs are used to produce vector hedgehogs and tensor ellipsoids. Glyphing is essentially a data-driven modeling operation. That is, a canonical geometric/topological structure is defined, and the structure is modified according to data values at a particular location. The location is typically a point in the dataset, but glyphs may be used to represent cells, datasets, or even the relationships between data. Figure 14 shows superquadric glyphs used to indicate position in the x - y plane. In a famous technique, Chernoff (1973) used the human face as a glyph, and linked data value to the features of the face. The key to glyph design is to use a representation that naturally conveys information to the viewer. This is inherently problem and application specific. Another issue with glyphs is that the use of too many clutters the display.

3.4.8 Other operations

A variety of other modeling operations are used in visualization. Many of these techniques are used to highlight data or provide various forms of visual annotation. Points and line primitives (streamlines, edges) are often thickened by placing spheres at points (another simple use of glyphing) or wrapping lines with tubes. Data may be framed or set

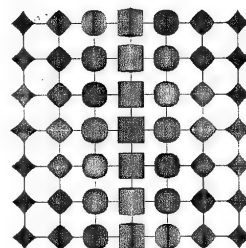


Figure 14. Glyphs used to indicate x - y location in the plane. Superquadrics are modified according to coordinate value. (Courtesy of Kivware, Inc. Taken from the book *The Visualization Toolkit: An Object-Oriented Approach to 3-D Graphics Third Edition* ISBN-1-530934-07-6.) A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

off by employing frames or base planes. Shadows are often used to provide depth cues. In some cases, extrusion is used to turn lower-dimensional data into more visually pleasing 3-D forms (carpet plots are one example of this).

Several computational geometry techniques are in common usage. Decimation is the process of reducing a mesh to a smaller size, preserving as much as possible the accuracy of the original mesh (Schroeder, Zarge and Lorensen, 1992). This is particularly important where large data size prevents interactive performance (see Section 1.5). Mesh smoothing using Laplacian vertex placement (Taubin, 1995) or windowed sinc functions (Taubin, Zhang and Golub, 1996) are used to smooth out surface noise. For example, 3-D isocontours produced from medical data often reflect aliasing due to sampling or noise inherent to the imaging modality. Smoothing can effectively reduce the high-frequency noise characteristic of these conditions. Another common class of operations is that based on connectivity. Region growing in images – especially when coupled with statistical measures such as mean and variance – can effectively segment data of interest. Geometric connectivity is useful for extracting meshes and leaving behind small artifacts due to noise.

4 VOLUME RENDERING

Early computer graphics focused on surface rendering techniques. Ray tracing (Whitted, 1980) was and still is a popular technique based on tracking the interaction of

light rays with the surfaces in a scene (the surface typically being defined with implicit functions, collections of linear primitives, or splines). Indeed, current graphics techniques remain surface oriented – polygons are used to represent surfaces (even splines such as NURBS are tessellated into triangles) that are then processed by graphics hardware. However, visualization datasets are typically three-dimensional in nature and carry important information interior to the boundary. This has given rise to volume rendering techniques (Figure 15).

4.1 Overview

The basic idea behind volume rendering is simple: for each pixel in the image, information interior to the data set (lying under the pixel) is composited in an ordered fashion by applying a transfer function to the data. The transfer function – similar to a lookup table – is a mapping of data value to color and transparency (i.e. alpha value). The transfer function may consist of a simple set of piecewise continuous linear ramps for each of the red, green, blue, and alpha (RGBA) components, or it may be a complex segmentation and include the effect of gradients or other features of the data. Most often volume rendering is applied to a 3-D image datasets (i.e. volumes), but recent techniques have extended volume rendering to irregular data. In some cases, a resampling or probing operation is applied to irregular data to produce a regular volumetric dataset. This has the advantage of rendering speed at the potential cost of missing important features in the data.

Early volume-rendering techniques were based on pure software implementations and were slow (several minutes per image). Using parallel methods and hardware acceleration, rendering speeds have been improved to tens of frames per second. Two basic approaches to volume rendering are common: image-order and object-order methods.

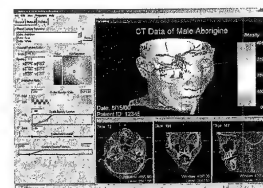


Figure 15. Volume rendering. (Image courtesy of VolView volume rendering system from Kivware, Inc.) A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

4.2 Image-order methods

Early versions of volume rendering used a ray-casting method. In this approach, rays are directed from the observer through each pixel in the output image (hence the name image-order) and into the 3-D data. Each ray is sampled in the order it propagates through the data, and at each sample point the data value is mapped through the transfer function to generate an RGBA color tuple that is blended with the previous color. Once the opacity reaches 1.0 (opaque), the ray is terminated; or if the ray reaches the far clipping plane, it is composited with the background color.

Ray-casting methods tend to be slower than the object-order methods described in the next section. However, ray-casting offers great flexibility and accuracy since the sampling rate and transfer function are easily adjusted based on characteristics of the data. The method is easily parallelized by either assigning different processors to different regions of the output image, or by dividing the data into pieces, rendering each piece separately, and compositing to form the output image. One important issue is the way in which sample points are interpolated from the surrounding data. The best results occur using tri linear interpolation from the eight voxel values surrounding any given sample. A faster approach used nearest-neighbor interpolation, which tends to produce heavily aliased results.

4.3 Object-order methods

While image-order methods start from the pixels in the output image, object-order methods start with the data (the object) and project it onto the view plane. Typically, data samples are projected onto the image plane in the form of an extended splat. In parallel (orthographic) projection, the splat is taken as a Gaussian kernel. In perspective projection, the splat is approximated with an ellipsoidal distribution. The size of the splat must be carefully controlled – too large a splat and the image is blurry, too small and gaps between adjacent splats appear.

Object-order methods are becoming popular because of the speed advantage they offer. Advances in graphics hardware – especially those relating to 2-D and 3-D texture – provide inexpensive, fast volume rendering solutions. Two-dimensional texture-based methods sample the volume with a series of planes, each plane having a texture mapped onto it (the texture includes transparency). If the planes are orthogonal to the x - y - z axes, artifacts are apparent in off-axis view directions. A better approach is to generate a series of planes perpendicular to the view direction. To produce the final image, the planes are rendered in order (e.g.

back to front) and each plane is blended into the current image. Emerging 3-D texture mapping provides a similar capability except that the interpolation of the texture onto the planes is performed in the graphics hardware.

5 METHODS IN LARGE DATA VISUALIZATION

A major goal of visualization is to assist the user in understanding large and complex data sets. However, as the size of analyses grows, it becomes more difficult to realize this goal, especially if what one desires is interactive data exploration. Several approaches are in use today as described in the following. However, this topic remains an area of active research and new methods are constantly being developed.

5.1 Culling

Probably the simplest approach to treating large data is to avoid processing/rendering data that is not visible. For example, a zoomed view of a scene may contain only a small fraction of the total data. Often, data is occluded by other surfaces. In these and other situations, culling algorithms can be used to render only those data that are visible. Such algorithms are typically based on bounding box, oriented bounding box, or bounding sphere culling. In some applications, the z -buffer may be used to eliminate data that is not visible. Visibility preprocessing can be used when the observer's viewpoint is known a priori. For example, views inside of buildings or other structures where the traversal path is known, and where regions are naturally separated by rooms or similar features may lend themselves to culling.

5.2 LOD

Level-of-detail methods are common in computer graphics. The idea is to replace a detailed representation with one or more coarser representations. Depending on the distance from the viewpoint, the coarsening scheme used, and the desired interaction rate, one of these levels is selected and used in place of the original data. In ideal cases, the resulting image fidelity remains essentially unchanged with the benefit of greatly improved interactivity. Even in situations where the coarsening is obvious, the ability to navigate (or position a camera) in an interactive setting is greatly enhanced.

5.3 Multiresolution

A natural extension of discrete level-of-detail methods is to adjust the resolution of the representation in a continuous, or nearly continuous fashion. Examples include octree-based methods that adjust resolution based on dynamic error measures (i.e. typically based on screen error or global geometric error). Several isocontouring methods based on octree decompositions have been described in the literature (Wilhelms and Van Gelder, 1992). Progressive meshes – based on a series of ordered edge-collapses – are used to simplify triangular (Hoppe, 1996) and tetrahedral meshes. Such meshes can be reconstituted from coarser to finer levels by a sequence of edge-based geomorph operations. This is one of a large number of algorithms used to reduce large meshes or terrain height fields (Laebke *et al.*, 2003). Structured data has the advantage that techniques in signal processing can be readily adapted to compress and transmit data including wavelet decompositions and feature detection (Machiraju *et al.*, 2000). Visualization systems may also take advantage of inherent data representations found in adaptive multiresolution grids (Berger and Olinger, 1984).

5.4 Out-of-core methods

A simple but highly effective approach to large data visualization is to process data out of main memory in a preprocessing step to produce output data that can be readily visualized using standard techniques. Thus, a three-dimensional data set of memory size $O(n^3)$ can be reduced in size to $O(n^2)$ or $O(n^1)$ as the data is transformed from its raw form to a visual representation. For example, a 3-D dataset can be isosurfaced to produce a surface; or streamlines may be produced to produce a 1-D polyline. Isocontouring (Chiang, Silva and Schroeder, 1998), streamlines (Ueng, Sikorski and Ma, 1997), vortex cores (Kenwright, 1998), and flow separation/attachment lines (Kenwright and Haines, 1997) are examples of such methods.

5.5 Parallel methods and data streaming

In conjunction with algorithmic approaches, system architecture can be designed to accommodate the demands of large data visualization. Certainly, parallel methods are important, utilizing approaches based on shared memory or distributed processing. Similar to those used for system solution, parallel methods for visualization must be carefully designed to minimize communication between processes. Visualization systems introduce additional constraints into the mix due to requirements on rendering and

interaction. For example, large tiled display driven by multiple processors must be synchronized and data carefully distributed to each processor. Depending on the parallel rendering technique used, data may be processed on an image-order or object-order basis (with meaning similar to the previous discussion of volume rendering). In object-order technique, the appropriate rendering primitives are dispersed to the processor responsible for a given tile. In image-order techniques, each processor produces an intermediate image, which is then composited to form the final image.

Another important technique is to stream data through the visualization pipeline in several pieces (Law *et al.*, 1999). Each piece is processed (with appropriate boundary conditions) and the data is assembled at the conclusion of processing. The assembly may occur during rendering (e.g. in a tiled display device) or using a compositing or append operation. Another advantage of this approach is that the user does not need to depend on system paging and may operate in pieces that fit in main memory for a significant improvement in elapsed time.

6 TAXONOMY FOR DATA VISUALIZATION SYSTEMS

Visualization systems can be categorized into one of three types of software systems: toolkits, development environments, and applications. This section provides a brief overview of their capabilities. Note that some of these systems provide multiple levels of functionality and may provide toolkit capabilities along with the features of an application or development environment. Furthermore, software is evolving rapidly and the systems described here represent only a small number of potential systems.

A distinguishing feature of many visualization systems is that they are architected around the concept of *data flow*. Data flow can be viewed as a series of operations on, or transformations of data. Ultimately, the results of analysis must be processed into graphics primitives that are then displayed by the graphics system. Furthermore, since visualization often involves interactive exploration, visualization systems must be flexible enough to *map* data from one data form to another. The data flow pipeline (or visualization pipeline) provides a simple abstraction that supports plugging in new process objects (algorithms) – often in complex combination – to produce new visualizations or transformations of existing data. This abstraction is natural to most users, is flexible, and can be readily codified into environments that support the construction of complex data processing networks from interlocking components.

6.1 Toolkits

Toolkits are collections of software modules that are used by developers to build development environments or applications. Software libraries – usually FORTRAN or C-based – provide procedural interfaces to visualization and graphics functionality. Software libraries have been generally replaced with object-oriented toolkits supporting multiple data representations (i.e. data objects) that are operated on by algorithms (i.e. process objects). The libraries in general include callback mechanisms and interfaces to popular GUI-building packages such as X, Motif, Windows, Tk, Qt, FLTK, and wxWindows. Toolkits include a wide variety of functionality ranging from IO to multithreading and parallel computing to memory management. The strength of toolkits is that they enable low-level control to algorithms and data representations and are designed to work alongside of other toolkits such as the GUI systems mentioned previously. However, while toolkits are inherently flexible, they require programming expertise to use. Two popular toolkits are VTK (Schroeder, Martin and Lorensen, 2003) and Open Inventor (<http://oss.sgi.com/projects/inventor/>). In addition, OpenGL is a popular graphics standard upon which most current visualization systems are built.

6.1.1 OpenGL

OpenGL is a graphics library providing low-level access to graphics functionality. While it is feasible to use OpenGL to create visualization applications, it does not provide standard visualization functionality such as isocontouring. Thus, significant effort is required to produce useful applications in practice. Instead, the systems mentioned in the following use OpenGL's rendering and interaction capabilities and add higher-level functionality that greatly simplify the task of building useful software tools.

6.1.2 VTK

VTK is an open-source, object-oriented toolkit supporting 3-D graphics, visualization, volume rendering, and image processing (www.vtk.org) (Schroeder, Martin and Lorensen, 2003). VTK provides hundreds of algorithms in an integrated system framework from which advanced applications can be created. While VTK does not define a GUI, it provides an interface to OpenGL and other popular tools such as X11, Windows, Qt, wxWindows, and Tk. Furthermore, VTK provides a set of powerful 3-D widgets that support operations on data such as clipping, cutting, transformation, and geometry creation. VTK is portable across all popular operating systems including Windows, Unix, Linux, and Mac OSX. Implemented in C++, VTK supports

language bindings to several interpreted languages such as Python, Tcl, and Java.

6.1.3 Open Inventor

Open Inventor is an object-oriented 3-D toolkit offering a comprehensive solution to interactive graphics programming problems. It presents a programming model based on a 3-D scene database that simplifies graphics programming. It includes a rich set of objects such as cubes, polygons, text, materials, cameras, lights, trackballs, handle boxes, 3-D viewers, and editors. It is built on top of OpenGL and defines a popular data exchange file format. However, unlike VTK, it does not include any visualization algorithms such as isocontouring or vector visualization. These must be added by the user. Another solution is to use VTK's data processing pipeline in conjunction with the functionality of Open Inventor.

6.2 Development environments

Development environments differ from toolkits in that they provide a GUI-based framework for creating visualization applications – typically by assembling data flow networks (so-called visual programming). Figure 16 is an example of such a network from the OpenDX software system. In addition, these systems also provide tools for building GUI and packaging functionality into applications. These tools are superb for crafting niche applications addressing a particular visualization need. However, as the application becomes more complex, the development environment tends to become more of a hindrance rather than a help since the complexity of the visual programming mechanisms often interfere with the low-level control required in a professional software tool. Development environments require programming skills, even if the skills are as simple as visually connecting modules into data flow networks.

6.2.1 AVS

One of the first true visualization development environments, AVS, brought significant attention to the emerging visualization field in the late 1980s and early 1990s. AVS provides a dataflow editing environment and GUI creation/packaging utility so that applications can be created and deployed rapidly. AVS is portable across a variety of platforms and maintains an open repository of contributed modules. See www.avs.com for more information.

6.2.2 OpenDX

OpenDX began as a commercial product offered by IBM and known as IBM Visualization Data Explorer. OpenDX

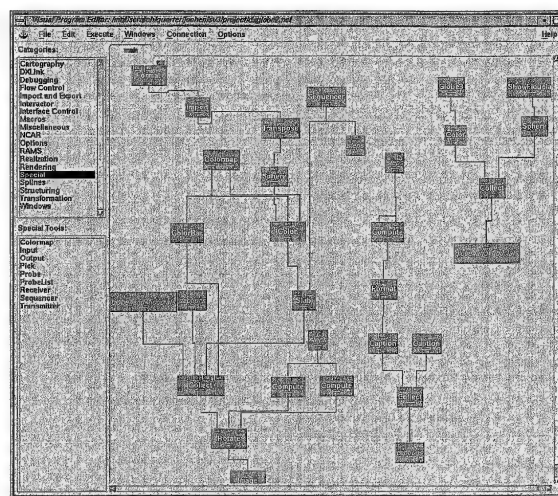


Figure 16. Typical data flow network as implemented in the OpenDX system. Courtesy of Kitware, Inc. Taken from the book *The Visualization Toolkit: An Object-Oriented Approach to 3-D Graphics* Third Edition ISBN-1-930934-07-6. A color version of this image is available at <http://www.mrw.interscience.wiley.com/lecm>

appeared in May 1999 when IBM converted DX into open-source software. OpenDX is known for its superlative visual programming environment and powerful data model. Development on OpenDX continues via a vital open-source community. Commercial support is also available for OpenDX from Visualization and Imagery Solutions (<http://www.vizsolutions.com/>). See <http://www.opendx.org> for more information about OpenDX.

6.2.3 TGS Amira

TGS provides a commercial development similar to OpenDX and AVS. Unlike these systems, Amira is available in application-specific bundles such as molecular visualization or virtual reality applications. See <http://www.tgs.com> for more information.

6.2.4 SCIRun

SCIRun is an extensive computational problem solving environment that includes powerful visualization capabilities. Unlike the previous systems, SCIRun integrates pre-processing, analysis, and visualization into an interactive environment for the solution of PDE's. Like the other development environments, visual programming is used to create data flow networks that link input, analysis, and visualization. SCIRun is designed to integrate with external packages and is extensible. SCIRun is available free for noncommercial use. Learn more about SCIRun from <http://software.sci.utah.edu/scirun.html>.

6.3 Applications

Turnkey applications require little or no programming. They are ideal for the engineer or scientist using visualization as

a tool to interact, manage, understand, and communicate about the computational process. Applications are generally the easiest to use if they support the capabilities required by the analyst. However, most applications are difficult to extend and become difficult to use if complex data exploration is required.

6.3.1 Intelligent light FieldView

FieldView is a commercial CFD postprocessor application supporting a variety of file formats, streamline generation, geometry/surface viewing, CFD calculator, probing, and presentation tools. See <http://www.ilight.com> for more information.

6.3.2 CEI EnSight Gold

CEI EnSight Gold is a general-purpose visualization application for analyzing, visualizing, and communicating high-end scientific and engineering datasets. EnSight Gold takes full advantage of parallel processing and rendering, provides support for an array of VR devices, and enables real-time collaboration. EnSight is used in automotive, biomedical, chemical processing, defense, and many other applications. See <http://www.ceintl.com/> for more information.

6.3.3 Kitware ParaView

ParaView is an open-source, turnkey application designed specifically for large data visualization employing scalable, distributed parallel processing (although it functions well on single processor systems). ParaView is built on top of VTK and is capable of exporting and importing VTK scripts. Designed as a general-purpose visualization tool, ParaView can be customized at run-time via Tcl scripts, or via plug-in XML modules defining a GUI associated with a data processing filter. ParaView supports several types of rendering, including hardware acceleration on a single processor when the data is small enough, or sort-last compositing on tiled displays if requested. See <http://www.paraview.org> for more information.

7 INTERFACING THE COMPUTATIONAL SYSTEM WITH THE VISUALIZATION SYSTEM

Often, the greatest obstacle facing the computational scientist wishing to use visualization is interfacing analysis solution data with the visualization system. Regular data

is generally easy to work with because the forms are simple – an image is an array of values in row or column major order. However, more complex analyses often use unstructured forms such as those found in finite element meshes. In terms of the data forms discussed in Section 1.2, the finite element mesh is irregular data that is attributed with various geometric, scalar, vector, and tensor information, and which can be visualized using the techniques discussed in the previous sections. The goal of this section is to address the technical issues when integrating irregular data in the form of finite element data with the visualization system. The specific topics covered include the mesh topological representation, determining the appropriate representation of information to be visualized on the mesh, and the transfer of mesh-based information to independent graphics structures.

7.1 Mesh topological representation

The classic finite element mesh representation defines an element in terms of an ordered list of nodes – with node point coordinates – that implicitly define both the topology and geometry of an element (e.g. a three-noded 2-D element defines a triangle whose nodes are at the vertices and the edges are straight lines; an eight-noded 2-D element defines a quadrilateral where four of the nodes define the vertices of the element and four are 'midside-nodes' for each of the four curved, quadratic edges). Such structures can be used to construct the information needed by visualization procedures by the proper interpretation of the ordered nodes, and when needed, construction of other topological information for the mesh. The need to support mesh representations that are adapted during simulation and the use of a broader set of finite element topology and shape combinations is leading to the use of richer and more formal representations. These representation of the mesh topology are in terms of the adjacency relationships between mesh regions, faces, edges, and vertices (see references Beall and Shephard (1997) and Remacle *et al.* (2002) for more information on such mesh data structures). Under the assumption that each topological mesh entity of dimension d , M^d , is bounded by a set of topological mesh entities of dimension $d-1$, M^{d-1} , the full set of mesh topological entities are

$$T_M = \{M\{M^0\}, M\{M^1\}, M\{M^2\}, M\{M^3\}\} \quad (12)$$

where $M\{M^d\}$, $d = 0, 1, 2, 3$ are respectively the set of vertices, edges, faces, and regions that define the primary topological elements of the mesh domain. It is possible to limit the mesh representation to just these entities under the following set of restrictions (Beall and Shephard, 1997):

1. Regions and faces have no interior holes.
2. Each entity of order d_i in a mesh, M^{d_i} , may use a particular entity of lower order, M^{d_j} , $d_j < d_i$, at most once.
3. For any entity M^{d_i} there is a unique set of entities of order $d_i - 1$, M^{d_i-1} that are on the boundary of M^{d_i} .

The first restriction means that regions may be directly represented by the faces that bound them, and faces may be represented by the edges that bound them. The second restriction allows the orientation of an entity to be defined in terms of its boundary entities (without the introduction of entity uses). For example, the orientation of an edge, M^1 , bounded by vertices M^0 and M^0 , is uniquely defined as going from M^0_j to M^0_k only if $j \neq k$. The third restriction means that a mesh entity is uniquely specified by its bounding entities.

As discussed in Beall and Shephard (1997), the use of more complete mesh topologies effectively supports general mesh modification operations and the ability to independently associate shape and other attribute information with the individual mesh entities. It also provides a convenient modeling abstraction when converting data from irregular form to the visualization system (see Chapter 17, this Volume).

7.2 Appropriate representation of information on the mesh

Often, the mesh-based geometric, scalar, vector, and tensor attribute information is not in a form suitable for processing by the visualization system. This mismatch is due to difference in the relative position at which information is stored (integration points versus element centers versus nodal points – positional mismatch). Furthermore, most visualization systems are limited in the element shape functions that they support (interpolation mismatch). As we saw previously, visualization systems typically support piecewise linear distributions defined in terms of the vertex values on simple 2-D or 3-D polyhedra. Interfacing with the visualization system means addressing these two forms of mismatch.

When the mismatch is due to interpolation, the simplest approach is to tessellate the element and produce a common form, typically linear simplices. Note that this mismatch can occur due to differences in either geometric interpolation (e.g. curvature of the mesh) or in solution interpolation (e.g. variation in solution unknowns). In principle, using the topological mesh hierarchy described previously, the mesh is tessellated first along edges, then faces, then regions, controlled by error measures due to shape and/or interpolation

variation. For example, an edge would be tessellated into piecewise linear line segments to minimize variation from the curved edge and to reduce solution error assuming linear interpolation along each line segment. A face using that edge as a boundary would begin with the edge tessellation as a boundary condition, and then tessellate the interior to satisfy the appropriate face error metrics. This process would continue through the element regions. The output is then a linear mesh.

When the mismatch is due to variation in the position of stored data, interpolation methods must be used. Direct and least squares projection procedures (Hinton and Campbell, 1974; Oden and Brauchi, 1971) are commonly applied for this purpose. Another simple approach is based on averaging. For example, data stored at element centers may be represented at vertices by averaging the data contained in all elements connected to that vertex. Note, however, that when using piecewise linear graphics primitives the quantities being displayed are C^0 and interpolation methods typically also assume that the data is C^0 on the mesh. There are situations in which the finite element information to be displayed is in fact C^{-1} (e.g. all variables when discontinuous Galerkin methods are used or when derivative quantities are being displayed in the case of C^0 Galerkin finite elements). However, there is no assurance that those fields are superior to the discontinuous ones. One way to address the basic accuracy of the field is to consider projection-based error estimation procedures (Blacker and Belytschko, 1994; Zienkiewicz and Zhu, 1992). As pointed out by Babuska *et al.* (1994), specific care must be taken with these procedures to ensure that they provide superior results at the evaluation points used. Another alternative is to apply various weak form (Remacle *et al.*, 2002) or variational (de Miranda and Ubertini, 2002) constructs for this process.

7.3 Transfer of solution data between computational grids

Several of the graphics techniques discussed previously operate most effectively on specific spatial structures such as regular samplings of data (e.g. volume rendering). Data may be sampled into another form to take advantage of these capabilities. In previous sections, this was called *probing*.

Any process that transforms information from one grid to another must determine the relationship of the cells in the sending grid to those in the receiving grid. Owing to the discrete nature of procedures employed, this means determining what cell in one grid contains a given point in the other grid and the parametric location relative to the interpolation of the solution in that element. Once the element is known, determination of the parametric location requires a

parametric inversion. Parametric inversion is trivial for linear and bilinear quadrilateral elements. However, for other element types whose shape functions are of higher order, it is not easily done in closed form. Methods for the iterative solution of nonlinear parametric inversions are given in Cheng (1988) and Crawford *et al.* (1989). Crawford *et al.* (1989) used elimination theory to reduce the system of equations to a single polynomial equation of higher order.

Naive methods to determine the target element by performing the parametric inversion of each element gives a computational growth rate of $O(n^3)$, which is clearly unacceptable. This can be addressed by the use of searching structures, typically based on various types of data trees, which yield a total searching time of $O(n \log n)$. These procedures provide a set of candidate elements to search that are in the local neighborhood of the point of interest. If the evaluation of these candidates determines that the point is not within any of those elements, procedures can be used to determine which element is the closest one to the point to use in the evaluation process (Niu and Shephard, 1990).

Data structures that can be used (Dannelongue and Tanguy, 1990; Zienkiewicz and Zhu, 1992) to determine a list of elements close to a given location include quadrees (Samet, 1990), K-d trees (Overmars and van Leeuwen, 1982), range trees (Samet, 1990), and alternating digital trees (Bonet and Peraire, 1991). Two interesting alternatives for general mesh to mesh transfers are range trees and alternating digital trees (ADT). The problem with these methods is the construction time (for balanced trees) and the inability to perform simple modifications to account for adaptive mesh modifications. In those cases in which the information is being evaluated onto a more uniform structure (e.g. uniform set of voxels or an octree), search methods that already employ such structures will be most advantageous. In fact, once a first point of interest is determined, the effective traversal of the mesh based on topological adjacency combined with traversal of the regular data structure can be made very efficient.

REFERENCES

- Barbuska I, Strouboulis T, Upadhyay CS and Gangaraj SK. Superconvergence in the finite element method by computer proof. *USACM Bull.* 1994; 7(3):10–25.
- Beall MW and Shephard MS. A general topology-based mesh data structure. *Int. J. Numer. Methods Eng.* 1997; 40(9):1573–1596.
- Berger M and Oliger J. Adaptive mesh refinement for hyperbolic partial differential equations. *J. Comput. Phys.* 1984; 53:484–512.
- Blackler TD and Belytschko T. Superconvergent patch recovery with equilibrium and conjoint interpolant enhancements. *Int. J. Numer. Methods Eng.* 1994; 37(3):517–536.
- Bonet J and Peraire J. An alternating digital tree (ADT) algorithm for 3d geometric searching and intersection problems. *Int. J. Numer. Methods Eng.* 1991; 31:1–17.
- Cheng JH. Automatic adaptive remeshing for finite element simulation of metal forming processes. *Int. J. Numer. Methods Eng.* 1988; 26:1–18.
- Chernoff H. Using faces to represent points in K -dimensional space graphically. *J. Am. Stat. Assoc.* 1973; 68:361–368.
- Chiang Y-J, Silva CT and Schroeder WJ. Interactive out-of-core isosurface extraction. *Proceedings of IEEE Visualization*. IEEE Computer Society Press: Los Alamitos, 1998.
- Conte SD and de Boor C. *Elementary Numerical Analysis*. McGraw-Hill, 1972.
- Crawford RH, Anderson DC and Waggenspack WN. Mesh rezoneing of 2-d isoparametric elements by inversion. *Int. J. Numer. Methods Eng.* 1989; 28:523–531.
- Dannelongue HH and Tanguy PA. Efficient data structures for adaptive remeshing with fem. *J. Comput. Phys.* 1990; 91:94–109.
- Delmarcelle T and Hesselink L. Visualizing second-order tensor fields with hyperstreamlines. *IEEE Comput. Graph. Appl.* 1993; 13(4):25–33.
- de Miranda S and Ubertini F. Recovery of consistent stresses for compatible finite elements. *Comput. Methods Appl. Mech. Eng.* 2002; 191:1595–1609.
- The First Information Visualization Symposium*. IEEE Computer Society Press: Los Alamitos, 1995.
- Foley JD, van Dam A, Fisher SK and Hughes JF. *Computer Graphics Principles and Practice* (2nd edn). Addison-Wesley: Reading, 1990.
- Globus A, Levit C and Lasinski T. A tool for visualizing the topology of three-dimensional vector fields. *Proceedings of Visualization '91*. IEEE Computer Society Press: Los Alamitos, 1991; 33–40.
- Helman JL and Hesselink L. Visualization of vector field topology in fluid flows. *IEEE Comput. Graph. Appl.* 1991; 11(3):36–46.
- Hinton E and Campbell JS. Local and Global smoothing of discontinuous finite element functions using a least squares method. *Int. J. Numer. Methods Eng.* 1974; 8:461–480.
- <http://loss.sgi.com/projects/inventor/>.
- Hoppe H. Progressive Meshes. *Comput. Graph. (Proc. SIGGRAPH '96)*.
- Kerwright DN. Automatic detection of open and closed separation and attachment lines. *Proceedings of Visualization '98*. IEEE Computer Society Press: Los Alamitos, 1998; 151–158.
- Kerwright DN and Haines R. Vortex identification - applications in aerodynamics: a case study. *Proceedings of Visualization '97*. IEEE Computer Society Press: Los Alamitos, 1997; 413–416.
- Law C, Martin KM, Schroeder WJ and Temkin JE. A multi-threaded streaming pipeline architecture for large structured data sets. In *Proceedings of IEEE Visualization '99*, October 1999; 225–232.
- Livnat Y, Shen Han-Wei and Johnson CR. A near optimal isosurface extraction algorithm using the span space. *IEEE Trans. Visualiz. Comput. Graph.* 1996; 2(1):73–84.
- Lorenson WE and Cline HE. Marching cubes: A high resolution 3D surface construction algorithm. *Comput. Graph. (Proc. SIGGRAPH)* 1987; 21(4):163–169.
- Luebke D, Martin R, Jonathan DC, Amitabh V, Benjamin W and Robert H. *Level of Detail for 3D Graphics*. Morgan Kaufmann, ISBN 1-55860-838-9, 2003.
- Machiraju R, Fowler JE, Thompson D, Schroeder WJ and Soti B. *EVTA: A Prototype System for Efficient Visualization and Interrogation of Terabyte Datasets*. Technical Report MSSU-COE-ERC-01-02, Engineering Research Center, Mississippi State University, 2000.
- McCormick BH, DeFanti TA and Brown MD. *Visualization in Scientific Computing*. Report of the NSF Advisory Panel on Graphics, Image Processing and Workstations, 1987.
- Nielsen GM and Hamman B. The asymptotic decider: resolving the ambiguity in marching cubes. In *Proceedings of IEEE Visualization '91*, San Diego, 1991.
- Niu Q and Shephard MS. *Transfer of Solution Variables Between Finite Element Meshes*. SCOREC Report 4-1990, Scientific Computation Research Center, RPI: Troy, New York, 1990.
- Oden JT and Brauchi HJ. On calculation of consistent stress distributions in finite element approximations. *Int. J. Numer. Methods Eng.* 1971; 4:337–357.
- Overmars M and van Leeuwen J. Dynamic multi-dimensional data structures based on quad and k-d trees. *Acta Inf.* 1982; 17(3):267–285.
- Remacle J-F, Klass O, Fleherty JE and Shephard MS. A parallel algorithm oriented mesh database. *Eng. Comput.* 2002; 18(3):274–284.
- Remacle J-F, Li X, Shephard MS and Chevaugnon N. Transient mesh adaptation using conforming and non-conforming mesh modifications. *11th International Meshing Roundtable*. Sandia National Laboratories, 2002; 261–272.
- Rosenblum L, Earnshaw RA, Encarnacao J and Hagen H. *Scientific Visualization Advances and Challenges*. Harcourt Brace & Company: London, 1994.
- Samet H. *The design and analysis of spatial data structures*. Addison-Wesley, 1990.
- Schroeder WJ, Martin KM and Lorenson WE. *The Visualization Toolkit: An Object Oriented Approach to 3D Graphics* (3rd edn). Kitware, Inc., ISBN 1-930934-07-6, 2003.
- Schroeder WJ, Volpe C and Lorenson WE. The stream polygon: a technique for 3D vector field visualization. *Proceedings of Visualization '91*. IEEE Computer Society Press: Los Alamitos, 1991; 126–132.
- Schroeder WJ, Zarge J and Lorenson WE. Decimation of triangle meshes. *Comput. Graph. (SIGGRAPH '92)* 1992; 26(2):65–70.
- Taubin G. A signal processing approach to fair surface design. *Comput. Graph. (Proc. SIGGRAPH)* 1995.
- Taubin G, Zhang T and Golub G. Optimal surface smoothing as filter design. *Fourth European Conference on Computer Vision (ECCV '96)*, Cambridge, UK, April 14–18, 1996, Proceedings, Volume I, Springer Verlag, 1996.
- Ueng SK, Sikorski K and Ma KL. Out-of-core streamline visualization on large unstructured meshes. *IEEE Trans. Visualiz. Comput. Graph.* 1997; 3(4):370–380.
- Watt A. *3D Computer Graphics* (2nd edn). Addison-Wesley: Reading, 1993.
- Whitted T. An improved illumination model for shaded display. *CACM* 1980; 23(6):343–349.
- Wilhelms J and Van Gelder A. Octrees for faster isosurface generation. *ACM Trans. Graph.* 1992; 11(3):201–227.
- Zienkiewicz OC and Zhu JZ. Superconvergent patch recovery and a posteriori error estimates, part 1: The recovery technique. *Int. J. Numer. Methods Eng.* 1992; 33(7):1331–1364.

Chapter 19

Linear Algebraic Solvers and Eigenvalue Analysis

Henk A. van der Vorst

Utrecht University, Utrecht, The Netherlands

| | |
|--|-----|
| 1 Introduction | 551 |
| 2 Mathematical Preliminaries | 551 |
| 3 Direct Methods for Linear Systems | 553 |
| 4 Preconditioning | 560 |
| 5 Incomplete LU Factorizations | 562 |
| 6 Methods for the Complete Eigenproblem | 567 |
| 7 Iterative Methods for the Eigenproblem | 571 |
| Notes | 574 |
| References | 575 |

1 INTRODUCTION

In this chapter, an overview of the most widely used numerical methods for the solution of linear systems of equations and for eigenproblems is presented.

For linear systems $Ax = b$, with A a real square nonsingular $n \times n$ matrix, direct solution methods and iterative methods are discussed. The direct methods are variations on Gaussian elimination. The iterative methods are the so-called Krylov projection-type methods, and they include popular methods such as Conjugate Gradients, MINRES, Bi-Conjugate Gradients, QMR, Bi-CGSTAB, and GMRES.

Iterative methods are often used in combination with the so-called preconditioning operators (easily invertible approximations for the operator of the system to be solved). We will give a brief overview of the various preconditioners that exist.

For the eigenproblems of the type $Ax = \lambda x$, the QR method, which is often considered to be a direct method

because of its very fast convergence, is discussed. Strictly speaking, there are no direct methods for the eigenproblem; all methods are necessarily iterative. The QR method is expensive for larger values of n , and for these larger values, a number of iterative methods, including the Lanczos method, Arnoldi's method, and the Jacobi–Davidson method are presented.

For a general background on linear algebra for numerical applications, see Golub and Van Loan (1996) and Stewart (1998). Modern iterative methods for linear systems are discussed in van der Vorst (2003). A basic introduction with simple software is presented in Barrett *et al.* (1994). A complete overview of algorithms for eigenproblems, including pointers to software, is given in Bai *et al.* (2000). Implementation aspects for high-performance computers are discussed in detail in Dongarra *et al.* (1998).

Some useful state-of-the-art papers have appeared; we mention papers on the history of iterative methods by Golub and van der Vorst (2000) and Saad and van der Vorst (2000). An overview on parallelizable aspects of sparse matrix techniques is presented in Duff and van der Vorst (1999). A state-of-the-art overview for preconditioners is presented in Benzi (2002).

The purpose of this chapter is to make the reader familiar with the ideas and the usage of iterative methods. We expect that guided with sufficient knowledge about the background of iterative methods, one will be able to make a proper choice for a particular class of problems. It will also provide guidance on how to tune these methods, in particular, for the selection or construction of effective preconditioners.

2 MATHEMATICAL PRELIMINARIES

In this section, some basic notions and notations on linear systems and eigenproblems have been collected.

2.1 Matrices and vectors

We will be concerned with linear systems $Ax = b$, where A is usually an $n \times n$ matrix:

$$A \in \mathbb{R}^{n \times n}$$

The elements of A will be denoted as a_{ij} . The vectors $x = (x_1, x_2, \dots, x_n)^T$ and b belong to the linear space \mathbb{R}^n . Sometimes we will admit complex matrices $A \in \mathbb{C}^{n \times n}$ and vectors $x, b \in \mathbb{C}^n$, but that will be explicitly mentioned.

Over the space \mathbb{R}^n , we will use the Euclidean inner product between two vectors x and y :

$$x^T y = \sum_{i=1}^n x_i y_i$$

and for $v, w \in \mathbb{C}^n$, we use the standard complex inner product

$$v^H w = \sum_{i=1}^n \bar{v}_i w_i$$

These inner products lead to the 2-norm or Euclidean length of a vector

$$\begin{aligned} \|x\|_2 &= \sqrt{x^T x} \quad \text{for } x \in \mathbb{R}^n \\ \|v\|_2 &= \sqrt{v^H v} \quad \text{for } v \in \mathbb{C}^n \end{aligned}$$

With these norms, we can associate a 2-norm for matrices: for $A \in \mathbb{R}^{n \times n}$, its associated 2-norm $\|A\|_2$ is defined as

$$\|A\|_2 = \sup_{y \in \mathbb{R}^n, y \neq 0} \frac{\|Ay\|_2}{\|y\|_2}$$

and in the complex case, similarly, using the complex inner product. This matrix norm gives the maximal length multiplication effect of A on a vector (where the length is defined by the given norm).

The associated matrix norms are convenient because they can be used to bound products. For $A \in \mathbb{R}^{n \times k}$, $B \in \mathbb{R}^{k \times m}$, we have that

$$\|AB\|_2 \leq \|A\|_2 \|B\|_2$$

in particular,

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2$$

The inverse of a nonsingular matrix A is denoted as A^{-1} . Particularly useful is the condition number of a square nonsingular matrix A defined as

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2$$

The condition number is used to characterize the sensitivity of the solution x of $Ax = b$ with respect to perturbations in b and A . For perturbed systems, we have the following theorem.

Theorem 1 (Golub and Van Loan, 1996; Th. 2.7.2) Suppose

$$\begin{aligned} Ax &= b, \quad A \in \mathbb{R}^{n \times n}, \quad 0 \neq b \in \mathbb{R}^n \\ (A + \Delta A)y &= b + \Delta b, \quad \Delta A \in \mathbb{R}^{n \times n}, \quad \Delta b \in \mathbb{R}^n \end{aligned}$$

with $\|\Delta A\|_2 \leq \epsilon \|A\|_2$ and $\|\Delta b\|_2 \leq \epsilon \|b\|_2$.

If $\epsilon \kappa_2(A) = r < 1$, then $A + \Delta A$ is nonsingular and

$$\frac{\|y - x\|_2}{\|x\|_2} \leq \frac{\epsilon}{1-r} \kappa_2(A)$$

With the superscript T , we denote the transpose of a matrix (or vector): for $A \in \mathbb{R}^{n \times k}$, the matrix $B = A^T \in \mathbb{R}^{k \times n}$ is defined by

$$b_{ij} = a_{ji}$$

If $E \in \mathbb{C}^{n \times k}$, then the superscript H is used to denote its complex conjugate $F = E^H$, defined as

$$f_{ij} = \bar{e}_{ji}$$

Sometimes, the superscript T is used for complex matrices in order to denote the transpose of a complex matrix.

The matrix A is symmetric if $A = A^T$, and $B \in \mathbb{C}^{n \times n}$ is Hermitian if $B = B^H$. Hermitian matrices have the attractive property that their spectrum is real. In particular, Hermitian (or symmetric real) matrices that are positive-definite are attractive because they can be solved rather easily by proper iterative methods (the CG method).

A Hermitian matrix $A \in \mathbb{C}^{n \times n}$ is positive-definite if $x^H A x > 0$ for all $0 \neq x \in \mathbb{C}^n$. A positive-definite Hermitian matrix has only positive real eigenvalues.

We will encounter some special matrix forms, in particular tridiagonal matrices and (upper) Hessenberg matrices. The matrix $T = (t_{ij}) \in \mathbb{R}^{n \times n}$ will be called *tridiagonal*, if all elements for which $|i - j| > 1$ are zero. It is called *upper Hessenberg* if all elements for which $i > j + 1$ are zero. In the context of Krylov subspaces, these matrices are often $(k+1) \times k$ and they will then be denoted as $T_{k+1,k}$.

2.2 Eigenvalues and eigenvectors

For purposes of analysis, it is often helpful or instructive to transform a given matrix to an easier form, for instance, diagonal or upper triangular form.

The easiest situation is the symmetric case: for a real symmetric matrix, there exists an orthogonal matrix $Q \in \mathbb{R}^{n \times n}$, so that $Q^T A Q = D$, where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix. The diagonal elements of D are the eigenvalues of A , and the columns of Q are the corresponding eigenvectors of A . Note that the eigenvalues and eigenvectors of A are all real.

If $A \in \mathbb{C}^{n \times n}$ is Hermitian ($A = A^H$), then there exist $Q \in \mathbb{C}^{n \times n}$ and a diagonal matrix $D \in \mathbb{R}^{n \times n}$, so that $Q^H Q = I$ and $Q^H A Q = D$. This means that the eigenvalues of a Hermitian matrix are all real, but its eigenvectors may be complex.

Unsymmetric matrices do not, in general, have an orthonormal set of eigenvectors and may not have a complete set of eigenvectors, but they can be transformed unitarily to Schur form:

$$Q^* A Q = R$$

in which R is upper triangular.

If the matrix A is complex, then the matrices Q and R may be complex as well. However, they may be complex even when A is real unsymmetric. It may then be advantageous to work in real arithmetic. This can be realized because of the existence of the *real Schur decomposition*. If $A \in \mathbb{R}^{n \times n}$, then it can be transformed with an orthonormal $Q \in \mathbb{R}^{n \times n}$ as

$$Q^T A Q = \tilde{R}$$

with

$$\tilde{R} = \begin{bmatrix} \tilde{R}_{1,1} & \tilde{R}_{1,2} & \cdots & \tilde{R}_{1,k} \\ 0 & \tilde{R}_{2,2} & \cdots & \tilde{R}_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{R}_{k,k} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

Each $\tilde{R}_{i,j}$ is either 1×1 or a 2×2 (real) matrix having complex conjugate eigenvalues. For a proof of this, see Golub and Van Loan (1996, Chapter 7.4.1). This form of \tilde{R} is referred to as an upper *quasi-triangular matrix*.

If all eigenvalues are distinct, then there exists a nonsingular matrix X (in general not orthogonal) that transforms A to diagonal form:

$$X^{-1} A X = D$$

A general matrix can be transformed to Jordan form with a nonsingular X :

$$X^{-1} A X = \text{diag}(J_1, J_2, \dots, J_k)$$

where

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & & & \vdots \\ & & \ddots & & \\ \vdots & & & \ddots & \\ 0 & \cdots & 0 & & \lambda_i \end{bmatrix}$$

If there is a J_i with dimension greater than 1, then the matrix A is defective. In this case, A does not have a complete set of independent eigenvectors. In numerical computations, one may argue that small perturbations lead to different eigenvalues, and hence that it will be unlikely that A has a true Jordan form in actual computation. However, if A is close to a matrix with a nontrivial Jordan block, then this is reflected by a (severely) ill-conditioned eigenvector matrix X .

We will also encounter eigenvalues that are called *Ritz values*. For simplicity, we will introduce them here for the real case. The subspace methods that are collected in this chapter are based on the approach to identify good solutions from certain low-dimensional subspaces $\mathcal{V}^k \subset \mathbb{R}^n$, where $k \ll n$ denotes the dimension of the subspace. If $V_k \in \mathbb{R}^{n \times k}$ denotes an orthogonal basis of \mathcal{V}^k , then the operator $H_k = V_k^T A V_k \in \mathbb{R}^{k \times k}$ represents the projection of A onto V_k . Assume that the eigenvalues and eigenvectors of H_k are represented as

$$H_k s_j^{(k)} = \theta_j^{(k)} s_j^{(k)}$$

then $\theta_j^{(k)}$ is called a *Ritz value* of A with respect to \mathcal{V}^k and $V_k s_j^{(k)}$ is its corresponding *Ritz vector*. For a thorough discussion of Ritz values and Ritz vectors, see, for instance, Parlett (1980), Stewart (2001), and van der Vorst (2002).

For some methods, we will see that *Harmonic Ritz values* play a role. Let W_k denote an orthogonal basis for the subspace \mathcal{W}^k , then the Harmonic Ritz values of A with respect to that subspace are the inverses of the eigenvalues of the projection Z_k of A^{-1} :

$$Z_k = W_k^T A^{-1} W_k$$

3 DIRECT METHODS FOR LINEAR SYSTEMS

We will first consider the case that we have to solve

$$Ax = b$$

with A a square $n \times n$ matrix. The standard approaches are based upon Gaussian elimination. This works as follows.

Assuming that $a_{1,1} \neq 0$, one can subtract multiples of the first row of A of the other rows, so that the coefficients $a_{i,1}$ for $i > 1$ become zero. Of course, the same multiples of b_1 have to be subtracted from the corresponding b_i . This process can be repeated for the remaining $(n-1) \times (n-1)$ submatrix, in order to eliminate the coefficients for x_2 in the second column. After completion of the process, the remaining matrix has zeros below the diagonal and the linear system can now easily be solved. For a dense linear system, this way of computing the solution x requires roughly $(2/3)n^3$ arithmetic operations.

In order to make the process numerically stable, the rows of A are permuted so that the largest element (in absolute value) in the first column appears in the first position. This process is known as *partial pivoting* and it is repeated for the submatrices.

The process of Gaussian elimination is equivalent with the decomposition of A as

$$A = LU$$

with L a lower triangular matrix and U an upper triangular matrix and this is what is done in modern software. After the decomposition, one has to solve $LUx = b$ and this is done in two steps:

1. First solve y from $Ly = b$.
2. Then x is obtained from solving $Ux = y$.

The computational costs for one single linear system are exactly the same as for Gaussian elimination, and partial pivoting is included without noticeable costs. The permutations associated with the pivoting process are represented by an index array and this index array is used for rearranging b , before the solving of $Ly = b$.

If one has a number of linear systems with the same matrix A , but with different right-hand sides, then one can use the LU decomposition for all these right-hand sides. The solution for each new right-hand side then takes only $O(n^2)$ operations, which is much cheaper than to repeat the Gaussian elimination procedure afresh for each right-hand side.

This process of LU -decomposition, with partial pivoting, is the recommended strategy for the solution of dense linear systems. Reliable software for this process is available from software libraries including NAG and LAPACK (Anderson *et al.*, 1992) and the process is used in Matlab. It is relatively cheap to compute a good guess for the condition number of A , and this shows how sensitive the linear systems may be for perturbations to the elements of A and b (see Theorem 1). It should be noted that checking whether

the computed solution \hat{x} satisfies

$$\frac{\|b - A\hat{x}\|_2}{\|b\|_2} \leq \epsilon$$

for some small ϵ does not provide much information on the validity of \hat{x} without further information on A . If A is close to a singular matrix, then small changes in the input data (or even rounding errors) may lead to large errors in the computed solution. The condition number of A is a measure for how close A is to a singular matrix (cf. Theorem 1).

The computation of the factors L and U , with partial pivoting, is in general rather stable (small perturbations to A lead to acceptable perturbations in L and U). If this is a point of concern (visible through a relatively large residual $b - A\hat{x}$, then the effects of these perturbed L and U can be largely removed with *iterative refinement*. The idea of iterative refinement is to compute $r = b - A\hat{x}$ and to solve $Az = r$, using the available factors L and U . The computed solution \hat{z} is used to correct the approximated solution to $\hat{x} + \hat{z}$. The procedure can be repeated if necessary. Iterative refinement is most effective if r is computed in higher precision. Apart from this, the process is relatively cheap because it requires only n^2 operations (compared to the $O(n^3)$ operations for the LU factorization. For further details, we refer to Golub and Van Loan (1996).

For increasing n , the above sketched direct solution method becomes increasingly expensive ($O(n^3)$ arithmetic operations) and for that reason all sorts of alternative algorithms have been developed to help reduce the costs for special classes of systems.

An important subclass is the class of symmetric positive-definite matrices (see Section 2.1). A symmetric positive definite matrix A can be decomposed as

$$A = LL^T$$

and this is known as the *Cholesky* decomposition of A . The Cholesky decomposition can be computed in about half the time as an LU decomposition and pivoting is not necessary. It also requires half the amount of computer storage, since only half of A and only L need to be stored (L may even overwrite A if A is not necessary for other purposes). It may be good to note that the numerical stability of Cholesky's process does not automatically lead to accurate solutions x . This depends, again, on the condition number of the given matrix. The stability of the Cholesky process means that the computed factor L is relatively insensitive for perturbations of A .

There is an obvious way to transform $Ax = b$ into a system with a symmetric positive-definite matrix:

$$A^T Ax = A^T b$$

but this should be almost always avoided. It is not efficient because the construction of $B = A^T A$ requires $2n^3$ arithmetic operations for a dense $n \times n$ matrix. Moreover, the condition number of the matrix B is the square of the condition number of A , which makes the solution x much more sensitive to perturbations.

Another important class of matrices that occur in practical problems involves the matrices with many zero entries, the so-called sparse matrices. Depending on how the nonzero elements of A are distributed over the matrix, large savings can be achieved by taking account of the sparsity patterns. The easiest case is when the nonzero elements are in a (narrow) band around the diagonal of A . The LU factorization, even with partial pivoting, preserves much of this band structure, and software is available for these systems; see, for instance, LAPACK (Anderson *et al.*, 1992).

If the nonzero entries are not located in a narrow band along the diagonal, then it may be more problematic to take advantage of the given nonzero pattern (also called the *sparsity pattern*). It is often possible to permute rows and/or columns of A during the LU factorization process so that the factors L and U also remain satisfactorily sparse. It is not easy to code these algorithms, but software for the direct decomposition of sparse matrices is available (for instance in NAG).

For matrices with a very special sparsity pattern or where the elements satisfy special properties, for instance, constant diagonals as in Toeplitz matrices, special algorithms have been derived, for instance, Fast Poisson solvers and Toeplitz solvers. For an introduction and further references, see Golub and Van Loan (1996).

If the matrix A is nonsquare, that is, $m \times n$, or singular, then the Gaussian elimination procedures cannot be used. Instead, one may use QR factorizations, or even better (but more expensive), the singular value decomposition (SVD). For details on this, see Golub and Van Loan (1996) or the manuals of software libraries. The QR decomposition algorithm and the SVD are also available in Matlab.

The alternative for direct solvers, if any of the previously mentioned methods does not lead to the solution with reasonable computer resources (CPU time and storage), may be an iterative way of solution. Iterative methods are usually considered for the solution of very large sparse linear systems. Unfortunately, there is not one given iterative procedure that solves a linear system with a general sparse matrix, similar to the LU algorithm for dense linear systems. Iterative methods come in a great variety, and it requires much insight and tuning to adapt them for classes of special problems. Therefore,

we will pay more attention to these methods. This is necessary because many of the sparse problems, related to finite element or finite difference discretizations of mechanical problems, cannot be solved fast enough by direct methods.

3.1 Iterative solution methods

The idea behind iterative methods is to replace the given system by some nearby system that can be more easily solved; that is, instead of $Ax = b$, we solve the simpler system $Kx_0 = b$ and take x_0 as an approximation for x . Obviously, we want the correction z that satisfies

$$A(x_0 + z) = b$$

This leads to a new linear system

$$Az = b - Ax_0$$

Again, we solve this system by a nearby system, and most often one takes K again:

$$Kz_0 = b - Ax_0$$

This leads to the new approximation $x_1 = x_0 + z_0$. The correction procedure can now be repeated for x_1 , and so on, which gives us an iterative method.

For the basic or Richardson iteration, introduced above, it follows that

$$\begin{aligned} x_{k+1} &= x_k + z_k \\ &= x_k + K^{-1}(b - Ax_k) \\ &= x_k + K^{-1}r_k \end{aligned} \quad (1)$$

with $r_k = b - Ax_k$. We use K^{-1} only for notational purposes; we (almost) never compute inverses of matrices explicitly. When we speak of $K^{-1}b$, we mean the vector \tilde{b} that is solved from $K\tilde{b} = b$. The matrix K is called the *preconditioner*. In order to simplify our formulas, we will take $K = I$ and apply the presented iteration schemes to the preconditioned system $K^{-1}Ax = K^{-1}b$ if we have a better preconditioner available.

From now on, we will also assume that $x_0 = 0$ to simplify future formulas. This does not mean a loss of generality, because the situation $x_0 \neq 0$ can be transformed with a simple shift to the system

$$Ay = b - Ax_0 = \tilde{b} \quad (2)$$

for which obviously $y_0 = 0$.

For the simple Richardson iteration, it is easily shown that

$$x_k \in \text{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\} \quad (3)$$

$$\equiv K^k(A; r_0) \quad (4)$$

The k -dimensional space spanned by a given vector v , and increasing powers of A applied to v , up to the $(k-1)$ -th power, is called the k -dimensional Krylov subspace, generated with A and v , denoted by $K^k(A; v)$.

Apparently, the Richardson iteration, as it proceeds, delivers elements of Krylov subspaces of increasing dimension. It turns out that, at the expense of relatively little additional work compared to the Richardson method, we can identify much better approximations for the solution from the Krylov subspaces. This has led to the class of Krylov subspace methods that contain very effective iterative methods: Conjugate Gradients, GMRES, Bi-CGSTAB, and many more. Although older methods, such as SOR, can still be useful in certain circumstances, it is now generally accepted that the Krylov solvers, in combination with appropriate preconditioners, are the methods of choice for many sparse systems. For this reason, we restrict ourselves to a discussion of this class of methods. These methods are also very attractive because of their relation to iterative solvers for the eigenproblem. This means that in practice one can obtain, with relatively little extra costs, relevant information about the spectrum of A . This may guide the design of preconditioners, but it can also be used for safer stopping criteria and for sensitivity analysis (estimates for the condition number).

3.2 The Krylov subspace approach

Methods that attempt to generate better approximations from the Krylov subspace are often referred to as *Krylov subspace methods*. Because optimality usually refers to some sort of projection, they are also called *Krylov projection methods*. The Krylov subspace methods, for identifying suitable $x_k \in K^k(A; r_0)$, can be distinguished in four different classes (we will still assume that $x_0 = 0$):

1. The *Ritz-Galerkin approach*: Construct the x_k for which the residual is orthogonal to the current subspace: $b - Ax_k \perp K^k(A; r_0)$.
2. The *minimum norm residual approach*: Identify the x_k for which the Euclidean norm $\|b - Ax_k\|_2$ is minimal over $K^k(A; r_0)$.
3. The *Petrov-Galerkin approach*: Find an x_k so that the residual $b - Ax_k$ is orthogonal to some other suitable k -dimensional subspace.

4. The *minimum norm error approach*: Determine x_k in $A^T K^k(A^T; r_0)$ for which the Euclidean norm $\|x_k - x\|_2$ is minimal.

The Ritz-Galerkin approach leads to well-known methods such as Conjugate Gradients, the Lanczos method, the Full Orthogonalization Method (FOM), and Generalized Conjugate Gradients (GENCG). The minimum norm residual approach leads to methods like GMRES, MINRES, and ORTHODIR. The main disadvantage of these two approaches is that, for most unsymmetric systems, they lead to long and therefore expensive recurrence relations for the approximate solutions. This can be relieved by selecting other subspaces for the orthogonality condition (the Galerkin condition). If we select the k -dimensional subspace in the third approach as $K^k(A^T; r_0)$, then we obtain the Bi-CG and QMR methods, and these methods indeed work with short recurrences. The fourth approach is not so obvious, but for $A = A^T$, it leads to the SYMMLQ method of Paige and Saunders (1975).

Hybrids of these approaches have been proposed, like CGS, Bi-CGSTAB, Bi-CGSTAB(L), TFQMR, FGMRES, and GMRES.

The choice for a method is a delicate problem. If the matrix A is symmetric positive-definite, then the choice is easy: Conjugate Gradients. For other types of matrices, the situation is very diffuse. GMRES, proposed in 1986 by Saad and Schultz (1986), is the most robust method, but in terms of work per iteration step, it is also relatively expensive. Bi-CG, which was suggested by Fletcher (1976), is a relatively inexpensive alternative. The main disadvantage of Bi-CG is that it involves per iteration an operation with A and one with A^T . Bi-CGSTAB, proposed by van der Vorst (1992), is an efficient combination of Bi-CG and repeated 1-step GMRES, avoiding operations with A^T . Bi-CGSTAB requires about as many operations per iteration as Bi-CG. A more thorough discussion on Krylov methods is given in van der Vorst (2003). Other useful sources of information on iterative Krylov subspace methods include Axelsson (1994), Brezinski (1997), Bruaset (1995), Fischer (1996), Greenbaum (1997), Hackbusch (1994), Meurant (1999), and Saad (1996).

3.3 The Krylov subspace

In order to identify the approximations corresponding to the four different approaches, we need a suitable basis for the Krylov subspace.

Arnoldi (1951) proposed to compute an orthogonal basis as follows. Start with $v_1 \equiv r_0/\|r_0\|_2$. Then compute Av_1 , make it orthogonal to v_1 and normalize the result, which

```

v1 = r0 / ||r0||2
for j = 1, ..., m-1
  t = Avj
  for i = 1, ..., j
    hji = vj^T t
  end
  hjj = ||t||2
  vj+1 = t / hjj
end

```

Figure 1. Arnoldi's method with modified Gram-Schmidt orthogonalization.

gives v_2 . The general procedure is as follows: Assuming we already have an orthonormal basis v_1, \dots, v_j for $K^j(A; r_0)$, this basis is expanded by computing $t = Av_j$ and by orthonormalizing this vector t with respect to v_1, \dots, v_j .

This leads to an algorithm for the creation of an orthonormal basis for $K^m(A; r_0)$, as in Figure 1. The orthogonalization can be conveniently expressed in matrix terms. Let V_j denote the matrix with columns v_1 up to v_j , then it follows that

$$AV_{m-1} = V_m H_{m,m-1} \quad (5)$$

The $m \times (m-1)$ matrix $H_{m,m-1}$ is upper Hessenberg, and its elements $h_{i,j}$ are defined by the Arnoldi algorithm.

From a computational point of view, this construction is composed of three basic elements: a matrix-vector product with A , inner products, and vector updates. We see that this orthogonalization becomes increasingly expensive for increasing dimension of the subspace, since the computation of each $h_{i,j}$ requires an inner product and a vector update.

Note that if A is symmetric, then so is $H_{m-1,m-1} = V_{m-1}^T AV_{m-1}$, so that in this situation $H_{m-1,m-1}$ is tridiagonal. This means that in the orthogonalization process, each new vector has to be orthogonalized with respect to the previous two vectors only, since all other inner products vanish. The resulting three-term recurrence relation for the basis vectors of $K_m(A; r_0)$ is known as the *Lanczos method* (Lanczos, 1950) and some very elegant methods are derived from it. In this symmetric case, the orthogonalization process involves constant arithmetical costs per iteration step: one matrix-vector product, two inner products, and two vector updates.

We have discussed the construction of an orthonormal basis for the Krylov subspace because this also plays an important role for iterative eigenvalue solvers. Furthermore, the construction defines some matrices that play a role in the description of the iterative solvers. In practice, the user will never be concerned with the construction, because this is done automatically in all algorithms that will be presented.

3.4 The Ritz-Galerkin approach

The Ritz-Galerkin conditions imply that $r_k \perp K^k(A; r_0)$, and this is equivalent to

$$V_k^T (b - Ax_k) = 0$$

Since $b = r_0 = \|r_0\|_2 v_1$, it follows that $V_k^T b = \|r_0\|_2 e_1$ with e_1 the first canonical unit vector in \mathbb{R}^k . With $x_k = V_k y$ we obtain

$$V_k^T AV_k y = \|r_0\|_2 e_1$$

This system can be interpreted as the system $Ax = b$ projected onto the subspace $K^k(A; r_0)$.

Obviously, we have to construct the $k \times k$ matrix $V_k^T AV_k$, but this is immediately available from the orthogonalization process:

$$V_k^T AV_k = H_{k,k}$$

so that the x_k for which $r_k \perp K^k(A; r_0)$ can be easily computed by first solving $H_{k,k} y = \|r_0\|_2 e_1$, and then forming $x_k = V_k y$. This algorithm is known as FOM or GENCG (see Saad and Schultz, 1986).

When A is symmetric, then $H_{k,k}$ reduces to a tridiagonal matrix $T_{k,k}$, and the resulting method is known as the *Lanczos method* (Lanczos, 1952). When A is in addition positive-definite, then we obtain, at least formally, the *Conjugate Gradient* method. In commonly used implementations of this method, one implicitly forms an LU factorization for $T_{k,k}$, without generating $T_{k,k}$ itself, and this leads to very elegant short recurrences for the x_j and the corresponding r_j ; see the algorithm presented in Figure 2. This algorithm includes preconditioning with an operator

```

x0 is an initial guess, r0 = b - Ax0
for i = 1, 2, ...
  Solve Kw_{i-1} = r_{i-1}
  p_{i-1} = r_{i-1} / w_{i-1}
  if i = 1
    p_i = w_{i-1}
  else
    p_{i-1} = p_{i-1} / p_{i-2}
    p_i = w_{i-1} + p_{i-1} p_{i-2}
  end if
  q_i = Ap_i
  alpha_i = p_{i-1}^T q_i / p_{i-1}^T p_{i-1}
  x_i = x_{i-1} + alpha_i p_i
  r_i = r_{i-1} - alpha_i q_i
  if x_i accurate enough then quit
end

```

Figure 2. Conjugate gradients with preconditioning K .

K , which should be a fixed approximation for A throughout the entire iteration process.

The positive definiteness is necessary to guarantee the existence of the LU factorization, but it also guarantees that $\|x_k - x\|_A$ is minimal [1] over all possible x_k from the Krylov subspace of dimension k .

3.5 The minimum norm residual approach

We look for an $x_k \in K^k(A; r_0)$, that is, $x_k = V_k y$, for which $\|b - Ax_k\|_2$ is minimal. This norm can be rewritten, with $\rho = \|r_0\|_2$, as

$$\|b - Ax_k\|_2 = \|b - AV_k y\|_2 = \|\rho V_{k+1} e_1 - V_{k+1} H_{k+1,k} y\|_2$$

using the Krylov relation (5). Now we exploit the fact that V_{k+1} is an orthonormal transformation with respect to the Krylov subspace $K^{k+1}(A; r_0)$:

$$\|b - Ax_k\|_2 = \|\rho e_1 - H_{k+1,k} y\|_2$$

and this final norm can simply be minimized by solving the minimum norm least squares problem for the $(k+1) \times k$ matrix $H_{k+1,k}$ and right-hand side $\|r_0\|_2 e_1$. The least squares problem is solved by constructing a QR factorization of $H_{k+1,k}$, and because of the upper Hessenberg structure that can conveniently be done with Givens transformations (see Golub and Van Loan, 1996).

The GMRES method is based upon this approach. In order to avoid excessive storage requirements and computational costs for the orthogonalization, GMRES is usually restarted after each cycle of m iteration steps. This algorithm is referred to as GMRES(m); the not-restarted version is often called 'full' GMRES. There is no simple rule to determine a suitable value for m ; the speed of convergence over cycles of GMRES(m) may drastically vary for nearby values of m . It may be the case that GMRES($m+1$) is much more expensive than GMRES(m), even in terms of numbers of iterations.

We present in Figure 3, the modified Gram-Schmidt version of GMRES(m) for the solution of the linear system $Ax = b$. The application to preconditioned systems, for instance, $K^{-1}Ax = K^{-1}b$, is straightforward.

For an excellent overview of GMRES and related variants, such as FGMRES, see Saad (1996).

3.6 GMRES

There exist also variants of GMRES that permit for a variable preconditioning. This is particularly convenient

```

r = b - Ax_0, for a given initial guess x_0
k = x_0
for j = 1, 2, ...
  beta = ||r||_2, v_j = r / beta, beta = beta
  for i = 1, 2, ..., m
    w = Av_j
    h_{0,j} = v_j^T w, w = w - h_{0,j} v_j
    h_{1,j} = ||w||_2, v_{j+1} = w / h_{1,j}
    r_{1,j} = h_{1,j}
    for k = 2, ..., i
      gamma = c_{k-1} r_{k-1,j} + s_{k-1} h_{k,j}
      h_{k,j} = -s_{k-1} r_{k-1,j} + c_{k-1} h_{k,j}
      r_{k,j} = gamma
    delta = sqrt(r_{1,j}^2 + r_{i,j}^2), c_j = r_{1,j} / delta, s_j = h_{i+1,j} / delta
    r_{1,j} = c_j r_{1,j} + s_j h_{i+1,j}
    h_{i+1,j} = -s_j r_{1,j} + c_j h_{i+1,j}
    rho = delta / beta ( = ||b - Ax_{j+1,m}||_2 )
    if rho is small enough then
      (n_j = i, goto SOL)
  n_r = m, y_n_r = delta_n_r / r_{n_r}
SOL: for k = n_r - 1, ..., 1
  y_k = (delta_k - sum_{l=k+1}^{n_r} h_{l,k} y_l) / r_{k,k}
x = x_0 + sum_{j=1}^{n_r} y_j v_j, if rho small enough quit
r = b - Ax

```

Figure 3. Unpreconditioned GMRES(m).

in combination with domain decomposition methods, if the problem per domain is solved by an iterative solver itself.

The two most well-known variants are FGMRES (Saad, 1993) and GMRESR (van der Vorst and Vuk, 1994). Because GMRESR is, in the author's view, the most robust of the two, it has been presented here. The GMRESR algorithm can be described by the computational scheme in Figure 4.

```

x_0 is an initial guess; r_0 = b - Ax_0;
for i = 0, 1, 2, 3, ...
  Let z^{(m)} be the approximate solution of Ax = r_i
  obtained after m steps of an iterative method.
  c = Az^{(m)} (often available from the iterative method)
  for k = 0, ..., i-1
    alpha = (c_k, c)
    c = c - alpha c_k
  z^{(m)} = z^{(m)} - alpha c_k
end
c_j = c / ||c||_2; u_j = z^{(m)} / ||c||_2
x_{i+1} = x_i + (c_0, r_0) u_j
r_{i+1} = r_i - (c_0, r_0) c_j
if x_{i+1} is accurate enough then quit
end

```

Figure 4. The GMRESR algorithm.

A sufficient condition to avoid breakdown (when $\|c\|_2 = 0$) is that the norm of the residual at the end of an inner iteration is smaller than the right-hand residual: $\|Az^{(m)} - r_i\|_2 < \|r_i\|_2$. This can easily be controlled during the inner iteration process. If stagnation occurs, that is, no progress at all is made in the inner iteration, then van der Vorst and Vuk (1994) suggest doing one (or more) steps of the LSQR method, which guarantees a reduction (although this reduction is often only small).

When memory space is a limiting factor or when the computational costs per iteration become too high, we can simply truncate the algorithm (instead of restarting as in GMRES(m)). If we wish only to retain the last m vectors c_i and u_i , the truncation is effected by replacing the for k loop in Figure 4 by

$$\text{for } k = \max(0, i - m), \dots, i - 1$$

and of course, we have to adapt the remaining part of the algorithm so that only the last m vectors are kept in memory.

For a full discussion on GMRESR, see van der Vorst (2003).

3.7 The Petrov-Galerkin approach

For unsymmetric systems, we cannot, in general, reduce the matrix A to a tridiagonal system in a lower-dimensional subspace by orthogonal projections. The reason is that we cannot create an orthogonal basis for the Krylov subspace by a three-term recurrence relation (Faber and Manteuffel, 1984). We can, however, obtain a suitable nonorthogonal basis with a three-term recurrence, by requiring that this basis is orthogonal with respect to some other basis.

For this other basis, we select a convenient basis for the Krylov subspace generated with A^T and starting vector w_1 . It can be shown that a basis v_1, \dots, v_i for $K^i(A; v_1)$ can be created with a three-term recurrence relation, so that the v_j are orthogonal with respect to the w_k for $k \neq j$. The w_j are generated with the same recurrence relation as for the v_j , but with A replaced by A^T .

In matrix notation, this leads to $W_i^T A V_i = D_i T_{i,i}$, and also that $V_i^T A^T W_i = D_i^T T_{i,i}^T$, with $D_i = W_i^T V_i$ a diagonal matrix and $T_{i,i}$ a tridiagonal matrix. These bi-orthogonal sets of vectors form the basis for methods as Bi-CG and QMR.

Bi-CG is not as robust as CG. It may happen, for instance, that $w_i^T v_i = 0$ and then the method breaks down. Bi-CG is based on an LU decomposition of $T_{i,i}$, but since $T_{i,i}$ is not necessarily positive-definite or so, a flawless LU decomposition in bidiagonal L and U may not exist,

which gives another breakdown of the method. Fortunately, these circumstances do not occur frequently, but one has to be aware of them and carry out the required checks. There exist techniques to repair these breakdowns, but they require complicated coding. It may be convenient just to restart when a breakdown occurs, giving up some of the efficiency of the method and, of course, with a chance that breakdown occurs again. For a full treatment of Bi-CG, see van der Vorst (2003).

3.8 Bi-CGSTAB

Sonneveld showed that the two operations with A and A^T per iteration of Bi-CG can be replaced by two operations with A and with the effect that i iterations with Bi-CG are applied twice: once with the starting residual r_0 and then again with the same iteration constants on r_i . Surprisingly, this can be done for virtually the same computational costs as for Bi-CG, but the result is a method that often converges about twice as fast as Bi-CG. This method is known as CGS (Sonneveld, 1989).

Sonneveld's principle was further perfected in van der Vorst (1992) for the construction of Bi-CGSTAB in which the Bi-CG operations with A^T are replaced by operations with A and they are used to carry out GMRES(1) reductions on top of each Bi-CG iteration.

The preconditioned Bi-CGSTAB algorithm for solving the linear system $Ax = b$, with preconditioning K , reads as in Figure 5.

The matrix K in this scheme represents the preconditioning matrix and the way of preconditioning (van der Vorst, 1992). The above scheme, in fact, carries out the Bi-CGSTAB procedure for the explicitly preconditioned linear system

$$AK^{-1}y = b$$

but the vectors y_i and the residual have been back-transformed to the vectors x_i and r_i corresponding to the original system $Ax = b$.

The computational costs for Bi-CGSTAB are, per iteration, about the same as for Bi-CG. However, because of the additional GMRES(1) steps after each Bi-CG step, Bi-CGSTAB converges often considerably faster. Of course, Bi-CGSTAB may suffer from the same breakdown problems as Bi-CG. In an actual code, we should test for such situations and take appropriate measures, for example, restart with a different \tilde{r} ($= w_i$) or switch to another method (for example GMRES).

The method has been further generalized to Bi-CGSTAB(ℓ), which generates iterates that can be interpreted as the product of Bi-CG and repeated GMRES(ℓ).

```

 $x_0$  is an initial guess,  $r_0 = b - Ax_0$ 
Choose  $\tilde{r}$ , for example,  $\tilde{r} = r_0$ 
for  $i = 1, 2, \dots$ 
   $\rho_{i-1} = \tilde{r}^T r_{i-1}$ 
  if  $\rho_{i-1} = 0$  method fails
  if  $i = 1$ 
     $\beta_1 = 1$ 
  else
     $\beta_{i-1} = (\rho_{i-1} / \rho_{i-2}) (\alpha_{i-1} / \alpha_{i-2})$ 
     $\rho_i = \rho_{i-1} + \beta_{i-1} (\rho_{i-1} - \alpha_{i-1} \rho_{i-2})$ 
  endif
  Solve  $\hat{\rho}$  from  $K\hat{\rho} = \rho_i$ 
   $v_i = A\hat{\rho}$ 
   $\alpha_i = \rho_{i-1} / \tilde{r}^T v_i$ 
   $s = r_{i-1} - \alpha_i v_i$ 
  if  $\|s\|$  small enough then
     $x_i = x_{i-1} + \alpha_i \hat{\rho}$  quit
  Solve  $\hat{s}$  from  $K\hat{s} = s$ 
   $t = A\hat{s}$ 
   $\omega_i = t^T s / t^T t$ 
   $x_i = x_{i-1} + \alpha_i \hat{\rho} + \omega_i \hat{s}$ 
  if  $x_i$  is accurate enough then quit
   $r_i = s - \omega_i t$ 
  for continuation it is necessary that  $\omega_i \neq 0$ 
end

```

Figure 5. The Bi-CGSTAB algorithm with preconditioning.

For more details, see van der Vorst (2003). Software for this method is available from NAG.

4 PRECONDITIONING

4.1 Introduction

As we have seen in our discussions on the various Krylov subspace methods, they are not robust in the sense that they can be guaranteed to lead to acceptable approximate solutions within modest computing time and storage (modest with respect to alternative solution methods). For some methods (for instance, full GMRES), it is obvious that they lead, in exact arithmetic to the exact solution in maximal n iterations, but that may not be very practical. Other methods are restricted to specific classes of problems (CG, MINRES) or occasionally suffer from such nasty side-effects as stagnation or break down (Bi-CG, Bi-CGSTAB). Such poor convergence depends in a very complicated way on spectral properties (eigenvalue distribution, field of values, condition of the eigensystem, etc.) and this information is not available in practical situations.

The trick is then to try to find some nearby operator K such that $K^{-1}A$ has better (but still unknown) spectral properties. This is based on the observation that for $K = A$, we would have the ideal system $K^{-1}Ax = Ix = K^{-1}b$ and

all subspace methods would deliver the true solution in one single step. The hope is that for K in some sense close to A , a properly selected Krylov method applied to, for instance, $K^{-1}Ax = K^{-1}b$, would need only a few iterations to yield a good enough approximation for the solution of the given system $Ax = b$. An operator that is used with this purpose is called a *preconditioner* for the matrix A .

The general problem of finding an efficient preconditioner is to identify a linear operator K (the *preconditioner*) with the properties that [2]

- (1) K is a good approximation to A in some sense,
- (2) the cost of the construction of K is not prohibitive,
- (3) the system $Ky = z$ is much easier to solve than the original system.

There is a great freedom in the definition and construction of preconditioners for Krylov subspace methods. Note that in all the Krylov methods, one never needs to know individual elements of A , and one never has to modify parts of the given matrix. It is always sufficient to have a rule (subroutine) that generates, for given input vector y , the output vector z that can mathematically be described as $z = Ay$. This also holds for the nearby operator: it does not have to be an explicitly given matrix. However, one should realize that the operator (or subroutine) that generates the approximation for A can be mathematically represented as a matrix. It is then important to verify that application of the operator (or subroutine, or possibly even a complete code) on different inputs leads to outputs that have the same mathematical relation, with the same (possibly explicitly unknown) matrix K . For some methods, in particular Flexible GMRES and GMRESR, it is permitted that the operator K is (slightly) different for different input vectors (*variable preconditioning*). This plays an important role in the solution of nonlinear systems, if the Jacobian of the system is approximated by a Frechet derivative, and it is also attractive in some domain decomposition approaches (in particular, if the solution per domain itself is obtained by some iterative method again).

The following aspect is also important. One never (except for some trivial situations) forms the matrix $K^{-1}A$ explicitly. In many cases that would lead to a dense matrix and that would destroy all efficiency that could be obtained for the often sparse A . Even for dense matrix A , it might be too expensive to form the preconditioned matrix explicitly. Instead, for each required application of $K^{-1}A$ to some vector y , we first compute the result w of the operator A applied to y and then we determine the result z of the operator K^{-1} applied to w . This is often done by solving z from $Kz = w$, but there are also approaches by which approximations M for A^{-1} are constructed (e.g. *sparse approximate*

inverses) and then one applies, of course, the operator M to w in order to obtain z . Only very special and simple to invert preconditioners like diagonal matrices can be applied explicitly to A . This can be done before and in addition to the construction of another preconditioning.

Remember always that whatever preconditioner we construct, the goal is to reduce CPU time (or memory storage) for the computation of the desired approximated solution.

There are different ways of implementing preconditioning; for the same preconditioner, these different implementations lead to the same eigenvalues for the preconditioned matrices. However, the convergence behavior is also dependent on the eigenvectors or, more specifically, on the components of the starting residual in eigenvector directions. Since the different implementations can have quite different eigenvectors, we may thus expect that their convergence behavior can be quite different. Three different implementations are as follows:

1. **Left-preconditioning:** Apply the iterative method to $K^{-1}Ax = K^{-1}b$. We note that symmetry of A and K does not imply symmetry of $K^{-1}A$. However, if K is symmetric positive-definite, then $[x, y] = (x, Ky)$ defines a proper inner product. It is easy to verify that $K^{-1}A$ is symmetric with respect to the new inner product $[\cdot, \cdot]$, so that we can use methods like MINRES, SYMMLQ, and CG (when A is positive-definite as well) in this case. Popular formulations of preconditioned CG are based on this observation. If we are using a minimal norm residual method (GMRES or MINRES), we should note that with left-preconditioning, we are minimizing the preconditioned residual $K^{-1}(b - Ax)$, which may be quite different from the residual $b - Ax$. This could have consequences for stopping criteria that are based on the norm of the residual.
2. **Right-preconditioning:** Apply the iterative method to $AK^{-1}y = b$, with $x = K^{-1}y$. This form of preconditioning also does not lead to a symmetric product when A and K are symmetric. With right-preconditioning, we have to be careful with stopping criteria that are based upon the error: $\|y - y_k\|_2$ may be much smaller than the error-norm $\|x - x_k\|_2$ (equal to $\|K^{-1}(y - y_k)\|_2$) that we are interested in. Right-preconditioning has the advantage that it only affects the operator and not the right-hand side. This may be an attractive property in the design of software for specific applications.
3. **Two-sided preconditioning:** For a preconditioner K with $K = K_1 K_2$, the iterative method can be applied to $K_1^{-1} A K_2^{-1} z = K_1^{-1} b$, with $x = K_2^{-1} z$. This form of preconditioning may be used for preconditioners that come in factored form. This can be seen as a compromise between left- and right-preconditioning. This

form may be useful for obtaining a (near) symmetric operator for situations where K cannot be used for the definition of an inner product (as described under left-preconditioning).

Note that with all these forms of preconditioning, either explicit or implicit (through a redefinition of the inner product), we are generating a Krylov subspace for the preconditioned operator. This implies that the reduced matrix $H_{k,k}$ (cf. (5)) gives information about the preconditioned matrix; in particular, the Ritz values approximate eigenvalues of the preconditioned matrix. The generated Krylov subspace cannot be used in order to obtain information as well for the *unpreconditioned* matrix.

The choice of K varies from purely 'black box' algebraic techniques, which can be applied to general matrices, to 'problem-dependent' preconditioners that exploit special features of a particular problem class. Examples of the last class are discretized PDEs, where the preconditioner is constructed as the discretization of a nearby (easier to solve) PDE. Although problem-dependent preconditioners can be very powerful, there is still a practical need for efficient preconditioning techniques for large classes of problems.

There are only very few specialized cases where it is known a priori how to construct a good preconditioner and there are few proofs of convergence, except in very idealized cases. For a general system, however, the following approach may help to build up one's insight into what is happening. For a representative linear system, one starts with unpreconditioned GMRES(m), with m as high as possible. In one cycle of GMRES(m), the method explicitly constructs an upper Hessenberg matrix of order m , denoted by $H_{m,m}$. This matrix is reduced to upper triangular form, but before this takes place, one should compute the eigenvalues of $H_{m,m}$, called the *Ritz values*. These Ritz values usually give a fairly good impression of the most relevant parts of the spectrum of A . Then one does the same with the preconditioned system and inspects the effect on the spectrum. If there is no specific trend of improvement in the behavior of the Ritz values [3], when we try to improve the preconditioner, then obviously we have to look for another class of preconditioner. If there is a positive effect on the Ritz values, then this may give us some insight into how much more the preconditioner has to be improved in order to be effective. At all times, we have to keep in mind the rough analysis that we made in this chapter and check whether the construction of the preconditioner and its costs per iteration are still inexpensive enough to be amortized by an appropriate reduction in the number of iterations.

In this section, some of the more popular preconditioning techniques are described and references and pointers

for other techniques are given. The reader is referred to Axelsson (1994), Chan and van der Vorst (1997), Saad (1996), Meurant (1999), and van der Vorst (2003) for more complete overviews of (classes of) preconditioners. See Benzi (2002) for a very readable introduction to various concepts of preconditioning and for many references to specialized literature.

5 INCOMPLETE LU FACTORIZATIONS

Originally, preconditioners were based on direct solution methods in which part of the computation is skipped. This leads to the notion of *Incomplete LU* (or *ILU*) factorization (Meijerink and van der Vorst, 1977). We will now discuss these incomplete factorizations in more detail.

Standard Gaussian elimination is equivalent to factoring the matrix A as $A = LU$, where L is lower triangular and U is upper triangular. In actual computations, these factors are explicitly constructed. The main problem in sparse matrix computations is that the factors of A are often a good deal less sparse than A , which makes solving expensive. The basic idea in the point ILU preconditioner is to modify Gaussian elimination to allow fill-ins at only a restricted set of positions in the LU factors.

The following theorem collects results for the situation in which we determine a priori the positions of the elements that we wish to ignore during the Gaussian elimination process. Note that this is not a serious restriction because we may also neglect elements during the process according to certain criteria and this defines the positions implicitly. The indices of the elements to be ignored are collected in a set S :

$$S \subset S_n \equiv \{(i, j) \mid i \neq j, 1 \leq i \leq n, 1 \leq j \leq n\} \quad (6)$$

We can now formulate the theorem that guarantees the existence of incomplete decompositions for the M -matrix A (cf. Meijerink and van der Vorst, 1977, Th. 2.3).

Theorem 2. Let $A = (a_{ij})$ be an $n \times n$ M -matrix [4], then there exists for every $S \subset S_n$, a lower triangular matrix $\tilde{L} = (\tilde{l}_{ij})$, with $\tilde{l}_{ij} = 1$, an upper triangular matrix $\tilde{U} = (u_{ij})$, and a matrix $N = (n_{ij})$ with

- $\tilde{l}_{ij} = 0$, $u_{ij} = 0$, if $(i, j) \in S$
- $n_{ij} = 0$ if $(i, j) \notin S$, such that the splitting $A = \tilde{L}\tilde{U} - N$ leads to a convergent iteration (1).

The factors \tilde{L} and \tilde{U} are uniquely defined by S .

One can make, of course, variations on these incomplete splittings, for instance, by isolating the diagonal of \tilde{U} as a separate factor. When A is symmetric and positive-definite,

then it is obvious to select S so that it defines a symmetric sparsity pattern and then one can rewrite the factorization so that the diagonals of \tilde{L} and \tilde{U} are equal. This is known as an incomplete Cholesky decomposition.

A commonly used strategy is to define S by

$$S = \{(i, j) \mid a_{ij} \neq 0\} \quad (7)$$

That is, the only nonzeros allowed in the LU factors are those for which the corresponding entries in A are nonzero. It is easy to show that the elements k_{ij} of K match those of A on the set S :

$$k_{ij} = a_{ij} \quad \text{if } (i, j) \in S \quad (8)$$

Even though the condition (8) is sufficient (for certain classes of matrices) to determine the nonzero entries of L and U directly, it is more natural and simpler to compute these entries on the basis of a simple modification of the Gaussian elimination algorithm (see Figure 6). The main difference from the usual Gaussian elimination algorithm is in the innermost j -loop, where an update to a_{ij} is computed only if it is allowed by the constraint set S .

After the completion of the algorithm, the incomplete LU factors are stored in the corresponding lower and upper triangular parts of the array A . It can be shown that the computed LU factors satisfy (8).

The incomplete factors \tilde{L} and \tilde{U} define the preconditioner $K = (\tilde{L}\tilde{U})^{-1}$. In the context of an iterative solver, this means that we have to evaluate expressions like $z = (\tilde{L}\tilde{U})^{-1}y$ for any given vector y . This is done in two steps: first obtain w from the solution of $\tilde{L}w = y$ and then compute z from $\tilde{U}z = w$. Straightforward implementation of these processes leads to recursions, for which vector and parallel computers are not ideally suited. This sort of observation has led to reformulations of the preconditioner, for example, with reordering techniques or with blocking

```

ILU for an  $n \times n$  matrix  $A$  (cf. Axelsson, 1994):
for  $k = 1, 2, \dots, n-1$ 
   $d1/a_{kk}$ 
  for  $i = k+1, k+2, \dots, n$ 
    if  $(i, k) \in S$ 
       $e = d1/a_{kk}; a_{ik} = e$ 
      for  $j = k+1, \dots, n$ 
        if  $(i, j) \in S$  and  $(k, j) \in S$ 
           $a_{ij} = a_{ij} - ea_{kj}$ 
        end if
      end if
    end if
  end for
end for
end k

```

Figure 6. ILU for a general matrix A .

techniques. It has also led to different types of preconditioners, including diagonal scaling, polynomial preconditioning, and truncated Neumann series. These approaches may be useful in certain circumstances, but they tend to increase the computational complexity because they often lead to more iteration steps.

A well-known variant on ILU is the so-called *Modified ILU* (MILU) factorization (Dupont *et al.*, 1968; Gustafsson, 1978). For this variant, the condition (8) is replaced by

$$\sum_{j=1}^n k_{ij} = \sum_{j=1}^n a_{ij} + ch^2 \quad \text{for } i = 1, 2, \dots, n \quad (9)$$

The term ch^2 is for grid-oriented problems with mesh-size h . Although in many applications this term is skipped (that is, one often takes $c = 0$), this may lead to ineffective preconditioning (van der Vorst, 1989) or even breakdown of the preconditioner (see Eijkhout, 1992). In our context, the row sum requirement in (9) amounts to an additional correction to the diagonal entries compared to those computed in Figure 6. The correction leads to the observation that $Kz \approx Az$ for almost constant z (in fact, this was the motivation for the construction of these preconditioners). This results in very fast convergence for problems in which the solution is locally smooth. However, quite the opposite may be observed for problems in which the solution is far from smooth. For such problems, MILU may lead to much slower convergence than ILU.

The incomplete factorizations have been generalized with blocks of A instead of single elements. The inverses of diagonal blocks in these incomplete block factorizations are themselves again approximated, for instance, by their diagonal only or by the tridiagonal part (for details on this, see Axelsson, 1994; Meurant, 1999). In the author's experience, block incomplete decompositions can be quite effective for linear systems associated with two-dimensional PDEs, discretized over rectangular grids. However, for three-dimensional problems, they appeared to be less effective.

5.1 Reordering the unknowns

A standard trick for exploiting parallelism is to select all unknowns that have no direct relationship with each other and to number them first. For the 5-point finite difference discretization over rectangular grids, this approach is known as a *red-black ordering*. For elliptic PDEs, this leads to parallel preconditioners. The performance of the preconditioning step is as high as the performance of the matrix-vector product. However, changing the order of the

unknowns leads, in general, to a different preconditioner. Duff and Meurant (1989) report on experiments that show that most reordering schemes (for example, the red-black ordering) lead to a considerable increase in iteration steps (and hence in computing time) compared to the standard lexicographical ordering. For the red-black ordering, associated with the discretized Poisson equation, it can be shown that the condition number of the preconditioned system is only about one-quarter that of the unpreconditioned system for ILU, MILU, and SSOR, with no asymptotic improvement as the gridsize h tends to zero (Kuo and Chan, 1990).

One way to obtain a better balance between parallelism and fast convergence is to use more colors (Doi, 1991). In principle, since there is not necessarily any independence between different colors, using more colors decreases the parallelism but increases the global dependence and hence the convergence. In Doi and Hoshi (1992) up to 75 colors are used for a 76^2 grid on the NEC SX-3/14, resulting in a 2 Gflop/s performance, which is much better than for the wavefront ordering. With this large number of colors, the speed of convergence for the preconditioned process is virtually the same as with a lexicographical ordering (Doi, 1991).

The concept of *multicoloring* has been generalized to unstructured problems by Jones and Plassmann (1994). They propose effective heuristics for the identification of large independent subblocks of a given matrix. For problems large enough to get sufficient parallelism in these subblocks, their approach leads to impressive speedups compared to the natural ordering on a single processor.

Another approach, suggested by Meurant (1984), exploits the idea of the two-sided (or twisted) Gaussian elimination procedure for tridiagonal matrices. This is generalized for the incomplete factorization. By van der Vorst (1987), it is shown how this procedure can be done in a nested way. For 3D finite difference problems, twisting can be used for each dimension that gives an increase in parallelism by a factor of 2 per dimension. This leads, without further computational overhead, to incomplete decompositions, as well as triangular solves, that can be done in eight parallel parts (two in each dimension). For a discussion of these techniques, see Dongarra *et al.* (1998). This parallel ordering technique is sometimes referred to as 'vdi' ordering (Duff and Meurant, 1989) or 'van der Vorst' ordering (see, for example, Benzi, 2002).

A more sophisticated approach that combines ideas from twisting, domain decomposition with overlap, and reordering, was proposed by Magoumanga Made and van der Vorst (2001a,b, 2002). We will explain this idea for the special situation of a discretized second-order elliptic PDE over a

rectangular domain. The discretization has been carried out with the standard 5-point central difference stencil, which, over a rectangular grid with lexicographical ordering, leads to the familiar block matrix with 5 nonzero diagonals.

The first step is to split the domain in blocks, as in domain decomposition methods, and to order the unknowns lexicographically per block. This has been indicated, for the case of 8 horizontal blocks, in Figure 7. Per block, we start counting from one side ('the bottom layer'); the points on the last line ('the top layer') are ordered after all subdomains, as is indicated in Figure 8. For instance, the lines 1, 2, 3, and 26, all belong to the block stored with processor P_0 , but in the matrix interpretation, the first 3 lines are ordered first and line 26 appears in the matrix only after all other 'interior' lines. This means that the matrix has the following nonzero structure (we give only a relevant part of the matrix). Note that we have already introduced another element in our ordering, namely, the idea of twisting: the lines of the subdomains are ordered from bottom to top and from top to bottom in Figure 7.

Now imagine what happens if we carry out an incomplete LU factorization with zero fill. This would create level-1 fill in the error matrix. Note that, in particular, we would introduce fill in the subblock of the matrix that connects line 26 with line 5, and note also that we would not have seen this level-1 fill, if we would have selected all points lexicographically.

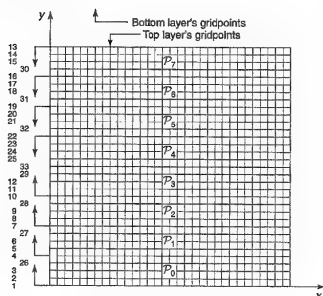


Figure 7. Decomposition of the grid into stripes, and assignment of subdomains to processors for $p = 8$. Arrows indicate the progressing direction of the line numbering per subdomain. Numbers along the y-axis give an example of global (line) ordering, which satisfies all the required conditions. Within each horizontal line, gridpoints are ordered lexicographically.

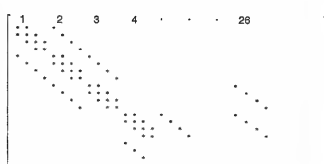


Figure 8. The structure of the reordered matrix.

This means that if we want the block ordering to be at least as effective as the standard ordering, we have to remove this additional fill. This can be interpreted as permitting level-1 fill in a small overlap, and this is the reason for the name *pseudo-overlap* for this way of ordering. How to generalize this idea for more arbitrary matrices is obvious: one compares the new ordering with the standard given one and one includes the possibly additional level-1 fill in the preconditioner. The idea can also be easily applied to preconditioners with a higher-level fill.

In Magoulou, Made and van der Vorst (2001a,b), it is suggested that the pseudo-overlap be increased and higher levels of fill that are introduced by the new block-wise ordering also be included. For high dimensional problems and relatively low numbers of processors, this leads to an almost negligible overhead. It is shown by analysis in Magoulou, Made and van der Vorst (2002) and by experiments mentioned in Magoulou, Made and van der Vorst (2001a,b) that the block ordering with pseudo-overlap may lead to parallelizable incomplete decompositions that are almost perfectly scalable if the number of processors p is less than \sqrt{n} , where n denotes the order of the given linear system (the reported experiments include experiments for 16 processors, for $n \approx 260\,000$).

5.2 Variants of ILU preconditioners

Many variants on the idea of incomplete or modified incomplete decomposition have been proposed in the literature. These variants are designed to reduce the total computational work, to improve the performance on vector or parallel computers, or to handle special problems. One could, for instance, think of incomplete variants of the various LU-decomposition algorithms discussed in Golub and Van Loan (1996, Chapter 4.4).

We will describe some of the more popular variants and give references to where more details can be found for other variants.

A natural approach is to allow more fill-in in the LU factor (that is a larger set S). Several possibilities have been proposed. The most obvious variant is to allow more fill-ins in specific locations in the \tilde{L} and \tilde{U} matrices (that is larger stencils) (see Axelsson and Barker, 1984; Gustafsson, 1978; Meijerink and van der Vorst, 1981). The most common location-based criterion is to allow a set number of levels of fill-in, where original entries have level zero, original zeros have level ∞ , and a fill-in in position (i, j) has level determined by

$$\text{Level}_{ij} = \min_{1 \leq k \leq \min(i, j)} \{\text{Level}_{ik} + \text{Level}_{kj} + 1\}$$

In the case of simple discretizations of partial differential equations, this gives a simple pattern for incomplete factorizations with different levels of fill-in. For example, if the matrix is from a 5-point discretization of the Laplacian in two dimensions, level 1 fill-in will give the original pattern plus a diagonal inside the outermost band (for instance, see Meijerink and van der Vorst, 1981).

The other main criterion for deciding which entries to omit is to replace the *drop-by-position* strategy by a *drop-by-size* one; that is, a fill-in entry is discarded if its absolute value is below a certain threshold value. This *drop-tolerance* strategy was first proposed by Munksgaard (1980). For the regular problems just mentioned, it is interesting that the level fill-in and drop strategies give a somewhat similar incomplete factorization because the numerical value of successive fill-in levels decreases markedly, reflecting the characteristic decay in the entries in the factors of the LU decomposition of A . For general problems, however, the two strategies can be significantly different. Since it is usually not known a priori how many entries will be above a selected threshold, the dropping strategy is normally combined with restricting the number of fill-ins allowed in each column (Saad, 1994). When using a threshold criterion, it is possible to change it dynamically during the factorization to attempt to achieve a target density of the factors (Axelsson and Munksgaard, 1983; Munksgaard, 1980; Saad, 1996) gives a very good overview of these techniques.

Although the notation is not yet fully standardized, the nomenclature commonly adopted for incomplete factorizations is $\text{ILU}(k)$, when k levels of fill-in are allowed, and $\text{ILUT}(\alpha, f)$ for the threshold criterion when entries of modulus less than α are dropped and the maximum number of fill-ins allowed in any column is f . There are many variations on these strategies and the criteria are sometimes

combined. In some cases, constraining the row sums of the incomplete factorization to match those of the matrix, as in MILU, can help (Gustafsson, 1978), but as we noted earlier, successful application of this technique is restricted to cases in which the solution of the (preconditioned) system is rather smooth.

Shifts can be introduced to prevent breakdown of the incomplete factorization process. As we have seen, incomplete decompositions exist for general M -matrices. It is well known that they may not exist if the matrix is positive-definite, but does not have the M -matrix property. Manteuffel (1980) considered incomplete Cholesky factorizations of diagonally shifted matrices. He proved that if A is symmetric positive-definite, then there exists a constant $\alpha > 0$, such that the Incomplete Cholesky factorization of $A + \alpha I$ exists. Since we make an incomplete factorization for $A + \alpha I$, instead of A , it is not necessarily the case that this factorization is also efficient as a preconditioner, the only purpose of the shift is to avoid breakdown of the decomposition process. Whether there exist suitable values for α such that the preconditioner exists and is efficient is a matter of trial and error.

Another point of concern is that for non- M -matrices the incomplete factors of A may be very ill conditioned. For instance, it has been demonstrated in van der Vorst (1981) that if A comes from a 5-point finite difference discretization of $\Delta u + \beta(u_x + u_y) = f$, then for sufficiently large β , the incomplete LU factors may be very ill conditioned even though A has a very modest condition number. Remedies for reducing the condition numbers of \tilde{L} and \tilde{U} have been discussed in Elman (1989) and van der Vorst (1981).

5.3 Hybrid techniques

In the classical incomplete decompositions, one ignores fill-in right from the start of the decomposition process. However, it might be a good idea to delay this until the matrix becomes too dense. This leads to a hybrid combination of direct and iterative techniques. One of such approaches has been described in Bomhof and van der Vorst (2000); we will describe it here in some detail.

We first permute the given matrix of the linear system $Ax = b$ to a doubly bordered block diagonal form:

$$\tilde{A} = P^T A P = \begin{bmatrix} A_{00} & 0 & \cdots & 0 & \cdots & A_{0n} \\ 0 & A_{11} & \ddots & \vdots & & A_{1n} \\ \vdots & \ddots & \ddots & 0 & & \vdots \\ 0 & \cdots & 0 & A_{n-1,n-1} & & \vdots \\ A_{n0} & A_{n1} & \cdots & \cdots & \cdots & A_{nn} \end{bmatrix} \quad (10)$$

Of course, the parallelism in the eventual method depends on the value of m , and some problems lend themselves more to this than others. Many circuit-simulation problems can be rewritten in an effective way, as a circuit is often composed of components that are only locally coupled to others.

We permute the right-hand side b as well to $\tilde{b} = P^T b$, which leads to the system

$$\tilde{A}\tilde{x} = \tilde{b} \quad (11)$$

with $x = Px$.

The parts of \tilde{b} and \tilde{x} that correspond to the block ordering will be denoted by \tilde{b}_i and \tilde{x}_i . The first step in the (parallelizable) algorithm will be to eliminate the unknown parts $\tilde{x}_0, \dots, \tilde{x}_{m-1}$, which is done by the Algorithm in Figure 9.

Note that S in Figure 9 denotes the Schur complement after the elimination of the blocks $0, 1, \dots, m-1$. In many relevant situations, direct solution of the reduced system $Sx_m = y_m$ requires the dominating part of the total computational costs, and this is where we bring in the iterative component of the algorithm.

The next step is to construct a preconditioner for the reduced system. This is based on discarding small elements in S . The elements larger than some threshold value define the preconditioner C :

$$c_{ij} = \begin{cases} s_{ij} & \text{if } |s_{ij}| > t|s_{ii}| \text{ or } |s_{ij}| > t|s_{jj}| \\ 0 & \text{elsewhere} \end{cases} \quad (12)$$

with a parameter $0 \leq t < 1$. In the experiments reported in Bomhof and van der Vorst (2000), the value $t = 0.02$ turned out to be satisfactory, but this may need some experimentation for specific problems.

When we take C as the preconditioner, then we have to solve systems like $Cv = w$, and this requires decomposition

```

Parallel_for i = 0, 1, ..., m-1
  Decompose  $A_i$ :  $A_i = L_i U_i$ 
   $L_m = A_m U_i^{-1}$ 
   $U_m = L_i^{-T} A_m$ 
   $y_i = L_i^{-T} b_i$ 
   $S_i = L_m U_{ij}$ 
   $z_i = L_m^{-1} y_i$ 
end
 $S = A_{mm} - \sum_{i=0}^{m-1} S_i$ 
 $y_m = b_m - \sum_{i=0}^{m-1} z_i$ 
Solve  $Sx_m = y_m$ 
Parallel_for i = 0, 1, ..., m-1
   $x_i = U_i^{-1} (y_i - U_{im}x_m)$ 
end

```

Figure 9. Parallel elimination.

of C . In order to prevent too much fill-in, reordering of C with a minimum degree ordering is suggested. The system $Sx_m = y_m$ is then solved with, for instance, GMRES with preconditioner C . For the examples described in Bomhof and van der Vorst (2000), it turns out that the convergence of GMRES was not very sensitive to the choice of t . The preconditioned iterative solution approach for the reduced system also offers opportunities for parallelism, although in Bomhof and van der Vorst (2000) it is shown that even in serial mode, the iterative solution (too sufficiently high precision) is often more efficient than direct solution of the reduced system.

In Bomhof and van der Vorst (2000), heuristics are described for the decision on when the switch from direct to iterative should take place. These heuristics are based on mild assumptions on the speed of convergence of GMRES. The paper also reports on a number of experiments for linear systems, not only from circuit simulation but also for some matrix problems taken from Matrix Market [5]. These experiments indicate that attractive savings in computational costs can be achieved, even in serial computation mode.

5.4 Element-by-element preconditioners

In finite element problems, it is not always possible or sensible to assemble the entire matrix, and it is as easy to form products of the matrix with vectors as when it is held in assembled form. Furthermore, it is easy to distribute such matrix multiplications to exploit parallelism. Hence, preconditioners are required that can be constructed at the element level. Hughes *et al.* (1983) were the first to propose such *element-by-element* preconditioners.

A parallel variant is suggested in Gustafsson and Lindslog (1986). For symmetric positive-definite A , they decompose each element matrix A_e as $A_e = L_e L_e^T$, and construct the preconditioner as $K = LL^T$, with

$$L = \sum_{e=1}^{n_e} L_e$$

In this approach, nonadjacent elements can be treated in parallel. An overview and discussion of parallel element-by-element preconditioners is given in van Gijzen (1994). To our knowledge, the effectiveness of element-by-element preconditioners is limited, in the sense that it does not often give a substantial improvement of the CPU time.

5.5 Preconditioning by blocks or domains

Other preconditioners that use direct methods are those where the direct method, or an incomplete version of it,

is used to solve a subproblem of the original problem. This can be done in *domain decomposition*, where problems on subdomains can be solved by a direct method, but the interaction between the subproblems is handled by an iterative technique.

Domain decomposition methods were motivated by parallel computing, but it now appears that the approach can also be used with success for the construction of global preconditioners. This is usually done for linear systems that arise from the discretization of a PDE. The idea is to split the given domain into subdomains, and to compute an approximation for the solution on each subdomain. If all connections between subdomains are ignored, this then leads to a *Block Jacobi* preconditioner. Chan and Goovaerts (1990) showed that the domain decomposition approach can actually lead to *improved* convergence rates, at least when the number of subdomains is not too large. This is because of the well-known divide-and-conquer effect when applied to methods with superlinear complexity such as ILU: it is more efficient to apply such methods to smaller problems and piece the global solution together.

In order to make the preconditioner more successful, one has to couple the domains, that is, one has to find proper boundary conditions along the interior boundaries of the subdomains. From a linear algebra point of view, this amounts to adapting the diagonal blocks in order to compensate for the neglected off-diagonal blocks. This is only successful if the matrix comes from a PDE problem and if certain smoothness conditions on the solution are assumed. If, for instance, the solution were constant, then one could remove the off-diagonal block entries, adding them to the diagonal block entries without changing the solution. Likewise, if the solution is assumed to be fairly smooth along a domain interface, one might expect this technique of diagonal block correction to be effective. Domain decomposition is used in an iterative fashion and usually the interior boundary conditions (in matrix language: the corrections to diagonal blocks) are based upon information from the approximate solutions on the neighboring subdomains that are available from a previous iteration step.

6 METHODS FOR THE COMPLETE EIGENPROBLEM

6.1 Introduction

Unlike the situation for linear systems solving, there are no truly direct methods for the solution of the eigenproblem, in the sense that, in general, one cannot compute the eigenvalues (or eigenvectors) exactly in a finite number of floating point operations. The iterative methods come in two different classes, one in which the matrix is driven to diagonal

form, or Schur form, by rotations, and one in which this is accomplished by detecting invariant subspaces. This second class of methods is explicitly based on the Power Method. The QR method is the most prominent member of this class; it converges so fast that the complete eigensystem of a matrix can be computed in a modest (but matrix dependent) factor times n^3 floating point operations. This is somewhat similar to the situation for the direct methods for solving dense linear systems. The QR method is, for this reason, often referred to as a direct method, in order to distinguish it from the essentially iterative subspace projection techniques. These iterative subspace projection techniques attempt to detect partial eigeninformation in much less than $O(n^3)$ work. They will be described in forthcoming sections.

Because the Power Method is important as a driving mechanism, although in hidden form, in the fast 'direct' methods as well as in the subspace projection methods, we will discuss it in some more detail. The reader should keep in mind, however, that the Power Method is seldom competitive as a stand-alone method; we need it here for a better understanding of the more superior techniques to come.

The Power Method is based on the observation that if we multiply a given vector v by the matrix A , then each eigenvector component in v is multiplied by the corresponding eigenvalue of A .

Assume that A is real symmetric, then it has real eigenvalues and a complete set of orthonormal eigenvectors

$$Ax_k = \lambda_k x_k, \quad \|x_k\|_2 = 1 \quad (k = 1, 2, \dots, n)$$

We further assume that the largest eigenvalue in modulus is single and that

$$|\lambda_1| > |\lambda_2| \geq \dots$$

Now, suppose we are given a vector v_1 , which can be expressed in terms of the eigenvectors as $v_1 = \sum_i \gamma_i x_i$, and we assume that $\gamma_1 \neq 0$ (that means that v_1 has a nonzero component in the direction of the eigenvector corresponding to λ_1).

Given this v_1 , we compute $Av_1, A(Av_1), \dots$, and it follows for the Rayleigh quotient of these vectors that

$$\lim_{j \rightarrow \infty} \frac{w_j^T A w_j}{w_j^T w_j} = \lambda_1$$

where $w_j = A^{j-1} v_1$.

The Power Method can be represented by the template given in Figure 10. The sequence $\theta^{(j)}$ converges, under the above assumptions, to the dominant (in absolute value)

```

v = v_i / ||v_i||_2
for i = 1, 2, ... until convergence
    v_{i+1} = Av
    θ^{(i)} = v^T v_{i+1}
    v = v_{i+1} / ||v_{i+1}||_2
end

```

Figure 10. The Power Method for symmetric A .

eigenvalue A . The scaling of the iteration vectors is necessary in order to prevent overflow or underflow. This scaling can be done a little bit cheaper by taking the maximum element of the vector instead of the norm. In that case, the maximum element of v_i converges to the largest (in absolute value) eigenvalue of A .

It is not hard to see why the Rayleigh quotients converge to the dominant eigenvalue. We first write w_j in terms of the eigenvectors of A :

$$w_j = \sum_{i \geq 1} \gamma_i \lambda_i^{j-1} x_i \\ = \lambda_1^{j-1} \left\{ \gamma_1 x_1 + \sum_{i \geq 2} \gamma_i \left(\frac{\lambda_i}{\lambda_1} \right)^{j-1} x_i \right\}$$

Hence,

$$\frac{w_j^T A w_j}{w_j^T w_j} = \lambda_1 \frac{\gamma_1^2 + \sum_{i \geq 2} \gamma_i^2 \left(\frac{\lambda_i}{\lambda_1} \right)^{2j-1}}{\gamma_1^2 + \sum_{i \geq 2} \gamma_i^2 \left(\frac{\lambda_i}{\lambda_1} \right)^{2j-2}} \\ = \lambda_1 \frac{\sum_{i \geq 1} \gamma_i^2 P_{2j-1}(\lambda_i)}{\sum_{i \geq 1} \gamma_i^2 P_{2j-2}(\lambda_i)}$$

with $P_j(t) = (t/\lambda_1)^j$.

For unsymmetric matrices, the situation is slightly more complicated, since A does not necessarily have an orthonormal set of eigenvectors. Let $A = XJX^{-1}$ denote the reduction to Jordan form, then $A^j v = XJ^j X^{-1}v$. If the largest eigenvalue, in modulus, is real and simple, then we see that this value will dominate, and by similar arguments as above, we see convergence in the direction of the corresponding column of X . If the largest eigenvalue is complex, and if A is real, then there is a conjugate eigenpair. If there is only one eigenpair of the same maximal modulus, then the vector $A^j v$ will ultimately have an oscillatory behavior and it can be shown (see Wilkinson, 1965, p. 579) that a combination of two successive vectors in the sequence $A^j v$ will converge to a subspace spanned by the two conjugate eigenvectors. The two eigenvectors can then be recovered by a least-squares solution approach. For matrices with elementary divisors, one can also show that the Power Method will

converge, although very slowly, to the eigenvector associated with the eigenvalue of the divisor.

The Power Method, as a 'stand-alone method,' received quite some attention until the early 1970s as the only means to extract eigenvalue information from large matrices. Early attempts to exploit the information hidden in a successive number of iterands, as, for instance, by Aitken acceleration, were only partially successful. In particular, the method of Lanczos (1950), which exploited all iterands of the Power Method, was a suspect for a long time, mainly because of its not-well-understood behavior in nonexact arithmetic. We will come back to this later.

If systems with the shifted matrix, like

$$(A - \sigma I)x = b$$

can be solved efficiently, then it is very profitable to use the Power Method with shifts. If one has a good guess σ for the eigenvalue that one is interested in, then one can apply the Power Method with $(A - \sigma I)^{-1}$. Note that it is not necessary to compute this inverted matrix explicitly. In the computation of the next vector $v_{j+1} = (A - \sigma I)^{-1} v_j$, the vector v_{j+1} can be solved from

$$(A - \sigma I)v_{j+1} = v_j$$

Assume that the largest eigenvalue (in absolute value) is λ_1 , and the second largest is λ_2 , and that σ is close to λ_1 . The speed of convergence now depends on the ratio

$$\frac{|\lambda_1 - \sigma|}{|\lambda_2 - \sigma|}$$

and this ratio may be a good deal smaller than $|\lambda_2/\lambda_1|$. Even when the solution of a shifted linear system is significantly more expensive than a matrix-vector multiplication with A , the much faster convergence may easily pay off for these additional costs.

We may update the shift as better approximations become available during the iteration process; that is, when we apply the algorithm in Figure 10 with $(A - \sigma)^{-1}$ at step i , then $\theta^{(i)}$ is an approximation for $1/(\lambda_1 - \sigma)$. This means that the approximation for λ_1 becomes $\sigma + (1/\theta^{(i)})$ and we can use this value as the shift for iteration $i+1$. This technique is known as *Rayleigh Quotient iteration*. Its convergence is ultimately cubic for symmetric matrices and quadratic for unsymmetric systems.

6.2 The QR method

We have already seen that for complex conjugate pairs, it is necessary to work with three successive vectors in the

Power iteration. This suggests that it may be a good idea to start with a block of vectors right from the start. So, let us assume that we start with a set of independent vectors $U_k^{(0)} = [u_1, u_2, \dots, u_k]$ and that we carry out the Power Method with $U_k^{(0)}$, which leads to the computation of

$$U_k^{(1)} = AU_k^{(0)}$$

per iteration. If we do this in a straightforward manner, then this will lead to unsatisfactory results because each of the columns of $U_k^{(0)}$ is effectively used as a starting vector for a single-vector Power Method, and all these single-vector processes will tend to converge toward the dominant eigenvector(s). This will make the columns of $U_k^{(1)}$ highly dependent in the course of the iteration. It is therefore a better idea to try to maintain better numerical independence between these columns and the most common technique for this is to make them orthonormal after each multiplication with A . This leads to the algorithm in Figure 11.

The columns of $U_k^{(1)}$ converge to a basis of an invariant subspace of dimension k under the assumption that the largest k (in absolute value) eigenvalues (counted according to multiplicity) are separated from the remainder of the spectrum. This can easily be seen from the same arguments as for the Power Method. If the eigenvalues are real, and the matrix is real, then the eigenvalues appear along the diagonal of R . For complex eigenvalues, the situation is slightly different; we will come back to this later.

In order to simplify the derivation of the QR method, we will first assume that A has real eigenvalues, which helps in avoiding complex arithmetic in the computation of the Schur forms.

We can apply the Orthogonal Iteration Method with a full set of starting vectors, say $U^{(0)} = I$, in order to try to reduce A to some convenient form. The matrix $U^{(0)}$ in the orthogonal subspace iteration converges to the matrix of Schur vectors, and the matrix $U^{(0)H}AU^{(0)}$ converges to Schur form. After one step of the algorithm: $AU^{(0)} = A = U^{(1)}R^{(1)}$, we can compute $A_1 = U^{(1)H}AU^{(1)}$, which is a similarity transform of the matrix A , and, hopefully, already a little bit more in the direction of Schur form than A itself.

```

start with orthonormal  $U_k^{(0)}$ 
for  $i = 1, \dots$  until convergence
     $V_k = AU_k^{(i-1)}$ 
    orthonormalize the columns of  $V_k$ :
     $V_k = Q_k R_k$ 
     $U_k^{(i)} = Q_k$ 
end

```

Figure 11. The Orthogonal Iteration Method.

Therefore, it might be a better idea to continue the algorithm with A_1 .

Let us define $Q_1 = U^{(1)}$. The matrix A_1 can be computed as

$$A_1 = Q_1^H A Q_1 = R^{(1)} Q_1$$

which simply reverses the factors of $A_0 = A$.

The next step in the Orthogonal Iteration Method is to compute AQ_1 and factor this as $U^{(2)}R^{(2)}$, but since $A = Q_1 R^{(1)}$, we see that

$$AQ_1 = Q_1 R^{(1)} Q_1$$

If we now factor $R^{(1)} Q_1$ as

$$R^{(1)} Q_1 = Q_2 R^{(2)} \quad (13)$$

then we see that

$$AQ_1 = U^{(2)} R^{(2)} \quad (14)$$

with $U^{(2)} = Q_1 Q_2$. Obviously, Q_2 gives the orthogonal transformation that corrects Q_1 to $U^{(2)}$. From equation (14), we see that

$$R^{(1)} Q_1 = A_1 = Q_1^H A Q_1 = Q_2 R^{(2)}$$

and hence the correction factor Q_2 would have also been obtained if we had continued the iteration with A_1 . This is a very nice observation because this correction factor drives the matrix A_1 again further to Schur form, and we can repeat this in an iterative manner.

The nice consequence of working with the transformed matrices A_i is that we can compute the correction matrix Q_{i+1} from the matrix product of the reverted factors Q_i and $R^{(i)}$ from the previous iteration, as we see from (13). This leads to the famous QR iteration.

The matrix A_i converges to Schur form, and the product $Q_1 Q_2 \dots Q_i$ of the correction matrices converges to the matrix of Schur vectors corresponding to A (remember that we are still assuming that all eigenvalues are real, in order to avoid complex arithmetic. We could have dropped this assumption, but then we should have used complex arithmetic; this can be avoided, however).

We can also apply the QR iteration with shifts, and this has very important consequences. Suppose we apply in one particular step of the algorithm a shift σ close to an eigenvalue λ_j ; that is, we do a QR factorization for $A_{i-1} - \sigma I$. If σ is close enough to an eigenvalue λ_j , then this implies that the matrix $A_{i-1} - \sigma I$, which is similar to $A - \sigma I$, is almost singular. This means that the matrix $R^{(i)}$ must be close to singular (since Q_i is orthogonal). If we do the

QR factorization with pivoting, in order to get the smallest diagonal element at the bottom position, then this means that one of the columns of Q_i must be close to the smallest singular vector of $A_{i-1} - \sigma I$, or better that that column is almost an eigenvector of A_{i-1} , and consequently, σ is almost an eigenvalue, and that is what we are looking for.

Pivoting is usually not necessary. If a subdiagonal entry of A_i is very small, then the eigenvalue computation can proceed for the two separate subblocks of A_i . If it is not small, then apparently the first $n-1$ columns of A_i are independent (recall that A_i is of upper Hessenberg form) so that (near) dependence must show up through a small diagonal entry at the bottom position in $R^{(i)}$.

The speed of convergence with which the smallest diagonal entry of $R^{(i)}$ goes to zero, for increasing i , is proportional to $|\lambda_{j+1} - \sigma|/|\lambda_j - \sigma|$, if λ_j is the eigenvalue closest to the shift and λ_{j+1} denotes the eigenvalue that is second closest. This is precisely the speed of convergence for the shift-and-invert Power Method and the surprising result is that we get the speed of shift and invert *without inverting anything*. A real shift can be easily incorporated in the QR algorithm. If we apply a shift in the i th iteration, then we have to QR-factor $A_{i-1} - \sigma I$, as

$$A_{i-1} - \sigma I \Rightarrow Q_i R^{(i)}$$

and, because of the orthogonality of Q_i , it follows that

$$Q_i^H (A_{i-1} - \sigma I) Q_i = R^{(i)} Q_i$$

However, in the next step, we want to continue with A_i (and possibly apply another shift), but that matrix is easily computed, since

$$A_i = Q_i^H A_{i-1} Q_i = R^{(i)} Q_i + \sigma I$$

This leads to the QR iteration with shifts, as given in Figure 12.

We have ignored some important aspects. One of these is that we have assumed that the real matrix, to which we applied the QR method, had real eigenvalues. This was necessary for the convergence of the matrix A_{i-1} (which is similar to A) to Schur form. For more general situations, that is, when the real matrix has conjugate pairs of eigenvalues (and eigenvectors), it can be shown that in real arithmetic the matrix converges to *generalized real Schur form*.

Recall that a matrix is in generalized real Schur form if it is block upper triangular and if the diagonal blocks are either 2×2 or one-dimensional. The 2×2 blocks along the diagonal of the generalized Schur form represent the complex conjugate eigenpairs of A (and the corresponding

```

start with  $A_0 = A$ 
for  $i = 1, \dots$ , until convergence
  factor  $A_{i-1} - \sigma I = QR^{(i)}$ 
  compute  $A_i = R^{(i)} Q_i + \sigma I$ 
end

```

Figure 12. The QR method with shifts σ .

eigenvectors can be computed by combining the corresponding columns of the accumulated transformation matrices Q_i).

This leaves the question how to apply complex shifts, since that would lead to complex arithmetic. It can be shown that by combining two successive steps in the QR algorithm, a pair of complex conjugate shifts can be used, without explicit complex arithmetic. A pair of such shifts is meaningful because complex eigenvalues of a real matrix always appear in conjugate pairs. If we want to accelerate convergence for one complex eigenpair, then clearly we also have to treat the conjugate member of this pair.

An important aspect of the QR algorithm is the computational complexity. If we apply it in the explained form, then it is quite expensive because we have to work with dense matrices in each iteration. Also, we have, for instance, to compute dense QR factorizations. These costs can be significantly reduced by first bringing the given matrix A , by orthogonal transformations, to upper Hessenberg form: $H = Q^H A Q$. Note that this is equivalent to the symmetric tridiagonal form if A is symmetric. The transformation to these forms can be done by Householder reflections or by Givens rotations (the Householder reflections are generally preferred since they are cheaper).

Another important observation is that the matrix Q_i can be easily computed for the shifted matrix, which leads to an implicit shift strategy.

For a detailed coverage of the QR method, see Golub and Van Loan (1996), Demmel (1997), and Watkins (1991, 1993, 2000).

The costs for the QR method, for computing all eigenvalues and eigenvectors, of a real unsymmetric matrix, are in the order of n^3 arithmetic operations. A crude and conservative upper estimate is $25n^3$ flops, including the initial reduction of A to upper Hessenberg form. This reduces to about $10n^3$ if only the eigenvalues are desired. In that case it is not necessary to accumulate the Q_i 's. In the symmetric case, the costs are reduced by roughly a factor of 2. These costs make the method only feasible for matrices on the order of a few thousands at most. For larger matrices, the method is used in combination with a technique that first reduces the large matrices to much smaller matrices (while attempting to preserve the main characteristics of the large matrix, in particular, the wanted eigenpairs).

Resuming: The QR method for eigenvalues and eigenvectors of a dense matrix is so effective because of the following ingredients:

1. Orthogonalization of the iterated columns in the basic orthogonal simultaneous iteration.
2. Accumulation of the orthogonal transformations, and their use as similarity transformations (this leads to the QR-RQ step).
3. The effect of shifts and their efficient implementation.
4. The initial reduction of A to upper Hessenberg (tridiagonal) form.

The QR method has been implemented in the major sub-routine libraries, in particular, in LAPACK, ScaLAPACK, IMSL, NAG, and MATLAB.

6.3 Generalized problems: the QZ method

The generalized eigenproblem

$$Ax = \lambda Bx \quad (15)$$

with $A, B \in \mathbb{C}^{n \times n}$, can be reduced to several canonical forms. Let us suppose that B is nonsingular, then we can define $y = Bx$, so that the eigenproblem (15) reduces to the standard eigenproblem

$$AB^{-1}y = \lambda y \quad (16)$$

However, working with AB^{-1} may be very unattractive, in particular when B is near singular. Also, the matrix AB^{-1} may be very nonnormal (or close to) defective. This makes the resulting eigenproblem highly sensitive to perturbations. Finally, with this approach, we cannot solve important classes of problems for which B is singular.

Moler and Stewart (1973) proved the interesting result that both A and B can be reduced to Schur form, albeit with two different unitary matrices Q and Z :

$$AZ = QR^A, \quad BZ = QR^B$$

with R^A and R^B upper triangular matrices. This leads to the so-called QZ method.

For a detailed description of the implementation of the QZ algorithm, see Moler and Stewart (1973) or Golub and Van Loan (1996, Chapter 7.7).

The total costs of the QZ algorithm, for the computation of the eigenvalues only, are about $30n^3$ flops. If one wants the eigenvectors as well, then the final Q and Z need to be evaluated (for the eigenvalues only, the products of all intermediate orthogonal transformations are not explicitly

necessary), which takes another $16n^3$ for Q and about $20n^3$ for Z (Golub and Van Loan, 1996, Chapter 7.7.7).

The (generalized) Schur form $R^A - \lambda R^B$ reveals the eigen structure of the original eigenproblem $A - \lambda B$. The pairs of diagonal elements of R^A and R^B define the eigenvalues λ . Here, we have to be careful; it is tempting to take the ratios of these diagonal elements, and, in view of the foregoing presentation, this is certainly correct. However, it sometimes hides sensitivity issues. Stewart (1978) has pointed at the asymmetry in the usual presentation of the generalized eigenproblem with respect to the role of A and B . Instead of the form $Ax - \lambda Bx = 0$, we might also have considered $\mu Ax - Bx = 0$, which is equivalent for $\mu = \lambda^{-1}$. If, for instance, B is singular, then the second form leads to the conclusion that there is an eigenvalue $\mu = 0$, which is equivalent to concluding that the first form has an eigenvalue $\lambda = \infty$. The important observation is that, in this case, in rounding arithmetic, there will be a tiny diagonal element r_{jj}^B in R^B , so that the corresponding ratio r_{jj}^A/r_{jj}^B leads to a very inaccurate eigenvalue approximation for λ , whereas the inverse leads to an accurate approximation for an eigenvalue μ .

The standard presentation for the generalized eigenproblem is quite appropriate for the type of problems that arise, for instance, in mechanical engineering, where A represents the *stiffness* matrix, with information of the differential operator, and B represents the *mass* matrix, which is largely determined by the chosen basis functions. For a well-chosen set of basis functions, the mass matrix is usually well-conditioned and then it is quite natural to consider the eigenvalue problem from the point of view of A .

7 ITERATIVE METHODS FOR THE EIGENPROBLEM

7.1 The Arnoldi method

With the Power method (see 6.1), we have generated the spanning vectors of a Krylov subspace

$$\mathcal{K}^m(A; v_1) = \text{span}\{v_1, Av_1, \dots, A^{m-1}v_1\}$$

However, note that the Power method exploits only the two most recently computed vectors for the computation of an approximating eigenvalue. The methods of Lanczos and Arnoldi exploit the whole Krylov subspace, and this leads not only to better approximations for the largest eigenvalue but also for many other eigenvalues.

From the definition of a Krylov subspace, it is clear that

$$\mathcal{K}^m(\alpha A + \beta I; v_1) = \mathcal{K}^m(A; v_1) \quad \text{for } \alpha \neq 0 \quad (17)$$

This says that the Krylov subspace is spanned by the same basis if A is scaled and/or shifted. The implication is that Krylov subspace methods for the spectrum of the matrix A are invariant under translations for A .

Krylov subspaces play a central role in iterative methods for eigenvalue computations. In order to identify better solutions in the Krylov subspace, we need a suitable basis for this subspace, one that can be extended inductively for subspaces of increasing dimension. The obvious basis $v_1, Av_1, \dots, A^{m-1}v_1$ for $K^m(A; v_1)$ is not very attractive from a numerical point of view, since the vectors $A^j v_1$ point more and more in the direction of the dominant eigenvalue for increasing j (the power method!), and hence the basis vectors become dependent in finite precision arithmetic.

Lanczos (1950) proposes to generate an orthogonal basis v_1, v_2, \dots, v_m for the Krylov subspace of dimension m , and shows that this could be done in a very economic way for symmetric A . Arnoldi (1951) describes a procedure for the computation of an orthonormal basis for unsymmetric matrices, and we have seen this procedure in Figure 1. The Lanczos algorithm follows as a special case.

The orthogonalization in Figure 1 leads to relations between the v_j that can be formulated in a compact algebraic form. Let V_j denote the matrix with columns v_1 up to v_j , $V_j = [v_1 | v_2 | \dots | v_j]$, then it follows that

$$AV_{m-1} = V_m H_{m,m-1} \quad (18)$$

The $m \times (m-1)$ matrix $H_{m,m-1}$ is upper Hessenberg, and its elements $h_{i,j}$ are defined by the Arnoldi orthogonalization algorithm. We will refer to this matrix $H_{m,m-1}$ as the *reduced matrix* A , or, more precisely, the matrix A reduced to the current Krylov subspace. From a computational point of view, this construction is composed from three basic elements: a matrix-vector product with A , inner products, and updates. We see that this orthogonalization becomes increasingly expensive for increasing dimension of the subspace, since the computation of each $h_{i,j}$ requires an inner product and a vector update.

7.2 The Ritz values and Ritz vectors

The eigenvalues of the leading $m \times m$ part $H_{m,m}$ of the matrix $H_{m+1,m}$ are called the *Ritz values* of A with respect to the Krylov subspace of dimension m . We will see later, in our discussions on the Lanczos and Arnoldi methods, how they can be related to eigenvalues of A . If s is an eigenvector of $H_{m,m}$, then $y = V_m s$ is called a *Ritz vector* of A . Also, these Ritz vectors can be related to eigenvectors of A .

All this is not so surprising because of the relation with the Krylov subspace and the vectors generated in the Power

Method. It seems obvious that we can base our expectations of the convergence of the Ritz values on our knowledge of convergence for the Power Method, something like, the largest Ritz value will converge faster to the largest eigenvalue of A than the Rayleigh quotient in the Power Method. But this kind of attitude is not very helpful for deeper insight. There is a very essential difference between Krylov subspace methods and the Power method and that is that the Krylov subspace is invariant for scaling of A and for shifting of A , that is, A and $\alpha A + \beta I$ generate exactly the same Krylov subspace. The reduced matrix $H_{m,m}$ is simply scaled and shifted to the reduced matrix $\tilde{H}_{m,m} = \alpha H_{m,m} + \beta I_m$, associated with $\tilde{A} = \alpha A + \beta I$. With I_m , we denote the $m \times m$ identity matrix.

This implies that for the Krylov subspace method, the notion of largest (in absolute value) eigenvalue loses its special meaning as opposed to the Power Method. Since the Krylov methods are shift invariant, the position of the origin in the spectrum is not relevant, and we rather make the distinction between *exterior* and *interior* eigenvalues. The Krylov methods have no special preference for eigenvalues that are at about the same distance from the center of the spectrum, provided that these eigenvalues are about equally well separated from the others. In particular, when the real spectrum of a symmetric real matrix is symmetrically distributed with respect to $(\lambda_1 + \lambda_n)/2$, λ_1 being the largest eigenvalue of A and λ_n the smallest one, then, for a starting vector that also has a symmetric weight distribution with respect to the corresponding eigenvectors, the convergence of the smallest Ritz value toward λ_n will be equally fast (or slow) as the convergence of the largest Ritz value to λ_1 .

For complex spectra, one has to consider the smallest circle that encloses all eigenvalues. With proper assumptions about the starting vector, one may expect that the eigenvalues close to this circle will be approximated fastest and that the more interior eigenvalues will be approximated later in the Krylov process (that is, for larger values of m). For more general starting vectors, this describes more or less the generic case, but with special starting vectors, one can force convergence toward favored parts of the spectrum.

The Arnoldi method forms the basis for the ARPACK software, described in Lehoucq *et al.* (1998). For more information on the method and for references, we refer to Bai *et al.* (2000).

7.3 The Lanczos method

Note that if A is symmetric, then so is

$$H_{m-1,m-1} = V_{m-1}^T A V_{m-1}$$

so that in this situation $H_{m-1,m-1}$ is tridiagonal:

$$AV_{m-1} = V_m T_{m,m-1} \quad (19)$$

The matrix $T_{m,m-1}$ is an $m \times (m-1)$ tridiagonal matrix, its leading $(m-1) \times (m-1)$ part is symmetric.

This means that in the orthogonalization process, each new vector has to be orthogonalized with respect to the previous two vectors only, since all other inner products vanish. The resulting three-term recurrence relation for the basis vectors of $K^m(A; v_1)$ is the kernel of the *Lanczos method* and some very elegant methods are derived from it. A template for the Lanczos method is given in the Algorithm shown in Figure 13.

In the symmetric case, the orthogonalization process involves constant arithmetical costs per iteration step: one matrix-vector product, two inner products, and two vector updates. In exact arithmetic, this process must terminate after at most n steps, since then the n orthonormal vectors for the Krylov subspace span the whole space. In fact, the process terminates after k steps if the starting vector has components only in the directions of eigenvectors corresponding to k different eigenvalues. In the case of finite termination, we have reduced the matrix A to triangular form with respect to an invariant subspace, and, in fact, the Lanczos' algorithm was initially viewed as a finite reduction algorithm for A .

For more information on the Lanczos method, we refer to Bai *et al.* (2000), which also gives pointers to software. The implementation of the Lanczos method is not trivial because of the effects of finite precision arithmetic that may lead to multiple copies of detected eigenpairs.

7.4 The two-sided Lanczos method: Bi-Lanczos

If the matrix A is Hermitian, then we can, at modest computational costs per iteration step, generate an orthonormal basis for the Krylov subspace $K^m(A; v_1)$. It suffices to put each newly generated vector v_j orthogonal to the two previous basis vectors v_j and v_{j-1} only. This leads to a reduced

```

v is a convenient starting vector
v1 = v / ||v||2, v0 = 0, beta0 = 0.
for i = 1, 2, ..., m-1
    t = Av_i - beta_{i-1}v_{i-1}
    alpha_i = v_i^T t
    t = t - alpha_i v_i
    beta_i = ||t||2
    v_{i+1} = t / beta_i
end

```

Figure 13. The Lanczos algorithm.

matrix that is symmetric tridiagonal. A theoretical result by Faber and Manteuffel (1984) shows that this is not possible for general unsymmetric matrices.

There exists a generalization of the Lanczos method for unsymmetric matrices. This method is known as the two-sided Lanczos method or the Bi-Lanczos method. There are several ways to derive this method; we follow the one in Wilkinson (1965, Chapter 6.36).

For ease of notation, we will restrict ourselves to the real case. For an unsymmetric matrix $A \in \mathbb{R}^{n \times n}$, we will try to obtain a suitable nonorthogonal basis with a three-term recurrence, by requiring that this basis is orthogonal with respect to some other basis.

We start from two Arnoldi-like recursions for the basis vectors of two Krylov subspaces, one with $A: K^m(A; v_1)$ and the other with $A^T: K^m(A^T; w_1)$. This leads to

$$h_{j+1,j} v_{j+1} = Av_j - \sum_{i=1}^j h_{i,j} v_i$$

$$g_{j+1,j} w_{j+1} = A^T w_j - \sum_{i=1}^j g_{i,j} w_i$$

and we require that v_{j+1} is orthogonal to all previous w_i and that w_{j+1} is orthogonal to all previous v_i . Clearly, this defines, apart from the constants $h_{j+1,j}$ and $g_{j+1,j}$, the vectors v_{j+1} and w_{j+1} , once the previous vectors are given. Then we have that

$$AV_j = V_{j+1} H_{j+1,j} \quad \text{and} \quad A^T W_j = W_{j+1} G_{j+1,j}$$

Since each new v_{j+1} is only orthogonal with respect to the w_i , for $i < j$, and likewise for w_{j+1} with respect to the v_i , it follows that

$$W_j^T V_j = L_{j,j} \quad \text{and} \quad V_j^T W_j = K_{j,j}$$

where $L_{j,j}$ and $K_{j,j}$ are lower triangular. Clearly,

$$K_{j,j}^T = L_{j,j}$$

so that both matrices are diagonal. Let us denote this matrix as D . Then, we have that

$$W_j^T A V_j = D H_{j,j}$$

and also

$$V_j^T A^T W_j = D G_{j,j}$$

This shows that $H_{j,j}$ and $G_{j,j}$ must be tridiagonal. The sets $\{v_j\}$ and $\{w_j\}$ are called biorthogonal.

```

Select a normalized pair  $v_1, w_1$  (for instance,  $w_1 = v_1$ )
such that  $w_1^T v_1 = \delta_1 \neq 0$ 
for  $i = 2, 3, \dots$ 
   $\rho = Av_i - \beta_{i-1}v_{i-1}$ 
   $\alpha_i = w_i^T \rho / \delta_i$ 
   $\rho = \rho - \alpha_i w_i$ 
   $\gamma_{i+1} = \|\rho\|_2$ 
   $v_{i+1} = \rho / \gamma_{i+1}$ 
   $w_{i+1} = (A^T w_i - \beta_{i-1}w_{i-1} - \alpha_i w_i) / \gamma_{i+1}$ 
   $\beta_i = w_{i+1}^T v_{i+1} / \delta_i$ 
   $\delta_i = \gamma_{i+1} \beta_i / \delta_{i-1}$ 
end.

```

Figure 14. The two-sided Lanczos algorithm.

In Figure 14, we show schematically an algorithm for the two-sided Lanczos method, suitable for execution with an unsymmetric matrix A .

In this algorithm, the vectors v_i are normalized: $\|v_i\|_2 = 1$, which defines automatically the scaling factors for the w_i . We have tacitly assumed that the inner products $w_i^T v_j \neq 0$, since the value 0 would lead to an unsolicited breakdown of the method if v_j and w_i are not equal to zero. If v_i is equal to zero, then we have an invariant subspace, and, in that case, the eigenvalues of T_i are eigenvalues of A .

For more information on this method and for pointers to software, we refer to Bai *et al.* (2000).

7.5 The Jacobi–Davidson method

The Jacobi–Davidson method is based on the projection of the given eigenproblem $Ax = \lambda x$ on a subspace that is in general not a Krylov subspace. The idea to work with non-Krylov subspaces was promoted in Davidson (1975). This leads to an approximate eigenpair, just as in the Arnoldi method. The main difference is the way in which the subspace is expanded. This is done by following an idea that was originally proposed by Jacobi (1846). Suppose that we have a normalized eigenvector approximation u_i with corresponding eigenvalue approximation θ_i , then the idea is to try to compute the correction t , so that

$$A(u_i + t) = \lambda(u_i + t)$$

for $t \perp u_i$ and λ close to θ_i . It can be shown that t satisfies the so-called correction equation

$$(I - u_i u_i^T)(A - \theta_i I)(I - u_i u_i^T)t = -(Au_i - \theta_i u_i)$$

This leads to the so-called Jacobi–Davidson method, proposed by Sleijpen and van der Vorst (1996).

```

Start with  $t = v_0$ , starting guess
for  $m = 1, \dots$ 
  for  $i = 1, \dots, m-1$ 
     $t = t - (v_i^T t) v_i$ 
   $v_m = t / \|t\|_2$ ,  $v_m^H = Av_m$ 
  for  $i = 1, \dots, m-1$ 
     $M_{im} = v_i^T v_m^H$ 
     $M_{mi} = v_m^T v_i^H$ 
   $M_{mm} = v_m^T v_m^H$ 
  Compute the largest eigenpair  $M_m = \theta_m$ 
  of the  $m$  by  $m$  matrix  $M$ , ( $\|e\|_2 = 1$ )
   $u = V e$  with  $V = [v_1, \dots, v_m]$ 
   $u^H = V^H e$  with  $V^H = [v_1^H, \dots, v_m^H]$ 
   $t = u^H - \theta_m u$ 
  If  $(\|t\|_2 \leq \epsilon)$ ,  $\lambda = \theta_m$ ,  $x = u$ , then STOP.
  Solve (approximately) a t.l.u. from
   $(I - uu^T)(A - \theta I)(I - uu^T)t = -r$ 

```

Figure 15. Jacobi–Davidson Algorithm for $\lambda_{\max}(A)$.

The basic form of this algorithm is given in Figure 15. In each iteration of this algorithm, an approximated eigenpair (θ, u) for the eigenpair of the matrix A , corresponding to the largest eigenvalue (in absolute value) of A , is computed. The iteration process is terminated as soon as the norm of the residual $Au - \theta u$ is below a given threshold ϵ .

To apply this algorithm, we need to specify a starting vector v_0 , and a tolerance ϵ . On completion, an approximation for the largest eigenvalue (in absolute value) $\lambda = \lambda_{\max}(A)$ and its corresponding eigenvector $x = x_{\max}$ is delivered. The computed eigenpair (λ, x) satisfies $\|Ax - \lambda x\| \leq \epsilon$.

The method is particularly attractive, that is fast converging, if one is able to solve the correction equation in some approximation at relatively low costs. For exact solution, one is rewarded with cubic convergence, but that may come at the price of an expensive solution of the correction equation. In practical applications, it is not necessary to solve this equation exactly. Effective solution requires the availability of a good preconditioner for the matrix $A - \theta_i I$. For more details on this, and for the efficient inclusion of the preconditioner, we refer to Bai *et al.* (2000).

NOTES

- [1] The A -norm is defined by $\|y\|_A^2 = (y, y)_A = (y, Ay)$.
- [2] The presentation in this chapter has partial overlap with Dongarra *et al.* (1998).
- [3] Ideally, the eigenvalues of $K^{-1}A$ should cluster around 1.
- [4] The nonsingular matrix A is called an M -matrix if all its off-diagonal elements are nonpositive and if all elements of A^{-1} are nonnegative.

- [5] Collection of testmatrices available at
<ftp://ftp.cise.ufl.edu/cis/tech-reports/tr98/tr98-016.ps>

REFERENCES

- Anderson E, Bai Z, Bischof C, Demmel J, Dongarra J, DuCroz J, Greenbaum A, Hammarling S, McKenney A, Ostouchov S and Sorensen D. *LAPACK User's Guide*. SIAM: Philadelphia, 1992.
- Arnoldi WE. The principle of minimized iteration in the solution of the matrix eigenproblem. *Q. Appl. Math.* 1951; 9:17–29.
- Axelsson O. *Iterative Solution Methods*. Cambridge University Press: Cambridge, 1994.
- Axelsson O and Barker VA. *Finite Element Solution of Boundary Value Problems: Theory and Computation*. Academic Press: New York, 1984.
- Axelsson O and Munksgaard N. Analysis of incomplete factorizations with fixed storage allocation. In *Preconditioning Methods – Theory and Applications*, Evans D (ed.). Gordon & Breach: New York, 1983; 265–293.
- Bai Z, Demmel J, Dongarra J, Ruhe A and van der Vorst H. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM: Philadelphia, 2000.
- Barrett R, Berry M, Chan T, Demmel J, Donato J, Dongarra J, Eijkhout V, Pozo R, Romine C and van der Vorst H. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM: Philadelphia, 1994.
- Benzi M. Preconditioning techniques for large linear systems: a survey. *J. Comput. Phys.* 2002; 182:418–477.
- Bombhof CW and van der Vorst HA. A parallel linear system solver for circuit-simulation problems. *Numer. Lin. Alg. Appl.* 2000; 7:649–665.
- Brezinski C. *Projection Methods for Systems of Equations*. North Holland: Amsterdam, 1997.
- Brusati AM. *A Survey of Preconditioned Iterative Methods*. Longman Scientific & Technical: Harlow, 1995.
- Chan TF and Goovaerts D. A note on the efficiency of domain decomposed incomplete factorizations. *SIAM J. Sci. Statist. Comput.* 1990; 11:794–803.
- Chan TF and van der Vorst HA. Approximate and incomplete factorizations. In *Parallel Numerical Algorithms: ICASE/LARC Interdisciplinary Series in Science and Engineering*, Keyes DE, Saneh A, Venkatakrishnan V (eds). Kluwer: Dordrecht, 1997; 167–202.
- Davidson ER. The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real symmetric matrices. *J. Comput. Phys.* 1975; 17:87–94.
- Demmel JW. *Applied Numerical Linear Algebra*. SIAM: Philadelphia, 1997.
- Doi S. On parallelism and convergence of incomplete LU factorizations. *Appl. Numer. Math.* 1991; 7:417–436.
- Doi S and Hoshi A. Large numbered multicolor MILU preconditioning on SX-3/14. *Int. J. Comput. Math.* 1992; 44:143–152.
- Dongarra JJ, Duff IS, Sorensen DC and van der Vorst HA. *Numerical Linear Algebra for High-Performance Computers*. SIAM: Philadelphia, 1998.
- Duff IS and Meurant GA. The effect of ordering on preconditioned conjugate gradient. *BIT* 1989; 29:635–657.
- Duff IS and van der Vorst HA. Developments and trends in the parallel solution of linear systems. *Parallel Comput.* 1999; 25:1931–1970.
- Dupont T, Kendall RP and Rachford HH Jr. An approximate factorization procedure for solving self-adjoint elliptic difference equations. *SIAM J. Numer. Anal.* 1968; 5(3):559–573.
- Eijkhout V. Beware of unperturbed modified incomplete point factorizations. In *Iterative Methods in Linear Algebra*, Beauwens R, de Groen P (eds). North Holland: Amsterdam, 1992; 583–591; IMACS Int. Symp., Brussels, 2–4 April 1991.
- Elman HC. Relaxed and stabilized incomplete factorizations for non-self-adjoint linear systems. *BIT* 1989; 29:890–915.
- Faber V and Manteuffel TA. Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM J. Numer. Anal.* 1984; 21(2):352–362.
- Fischer B. *Polynomial based iteration methods for symmetric linear systems*. Advances in Numerical Mathematics. Wiley and Teubner: Chichester, Stuttgart, 1996.
- Fletcher R. *Conjugate Gradient Methods for Indefinite Systems, Volume 506 of Lecture Notes Math.* Springer-Verlag: Berlin-Heidelberg-New York, 1976; 73–89.
- Golub GH and Van Loan CF. *Matrix Computations*. The Johns Hopkins University Press: Baltimore, 1996.
- Golub GH and van der Vorst HA. Numerical progress in eigenvalue computation in the 20th century. *J. Comput. Appl. Math.* 2000; 123(1–2):35–65.
- Greenbaum A. *Iterative Methods for Solving Linear Systems*. SIAM: Philadelphia, 1997.
- Gustafsson I. A class of first order factorization methods. *BIT* 1978; 18:142–156.
- Gustafsson I and Lindskog G. A preconditioning technique based on element matrix factorizations. *J. Comput. Methods Appl. Mech. Eng.* 1986; 55:201–220.
- Hackbusch W. *Iterative Solution of Large Sparse Systems of Equations*. Springer-Verlag: Berlin, 1994.
- Hughes TJR, Levit I and Winget J. An element-by-element solution algorithm for problems of structural and solid mechanics. *J. Comput. Methods Appl. Mech. Eng.* 1983; 36:241–254.
- Jacobi CGJ. Ueber ein leichtes Verfahren, die in der Theorie der Störungsstörungen vorkommenden Gleichungen numerisch aufzulösen. *J. Reine Angew. Math.* 1846; 30:51–94.
- Jones MT and Plassmann PE. The efficient parallel iterative solution of large sparse linear systems. In *Graph Theory and Sparse Matrix Computations*, IMA Vol. 56, George A, Gilbert JR, Liu JWH (eds). Springer-Verlag: Berlin, 1994; 229–245.
- Kuo JCC and Chan TF. Two-color Fourier analysis of iterative algorithms for elliptic problems with red/black ordering. *SIAM J. Sci. Statist. Comput.* 1990; 11:767–793.
- Lanczos C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Natl. Bur. Stand.* 1950; 45:225–280.

- Lanczos C. Solution of systems of linear equations by minimized iterations. *J. Res. Natl. Bur. Stand.* 1952; 49:33–53.
- Lehoucq RB, Sorensen DC and Yang C. *ARPACK User's Guide*. SIAM: Philadelphia, 1998.
- Magolu moaga Made M and van der Vorst HA. A generalized domain decomposition paradigm for parallel incomplete LU factorization preconditionings. *Future Gen. Comput. Syst.* 2001a; 17:925–932.
- Magolu moaga Made M and van der Vorst HA. Parallel incomplete factorizations with pseudo-overlapped subdomains. *Parallel Comput.* 2001b; 27:989–1008.
- Magolu moaga Made M and van der Vorst HA. Spectral analysis of parallel incomplete factorizations with implicit pseudo-overlap. *Numer. Lin. Alg. Appl.* 2002; 9:45–64.
- Manteuffel TA. An incomplete factorization technique for positive definite linear systems. *Math. Comp.* 1980; 31:473–497.
- Meijerink JA and van der Vorst HA. An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. *Math. Comp.* 1977; 31:148–162.
- Meijerink JA and van der Vorst HA. Guidelines for the usage of incomplete decompositions in solving sets of linear equations as they occur in practical problems. *J. Comput. Phys.* 1981; 44:134–155.
- Meurant G. *Numerical Experiments for the Preconditioned Conjugate Gradient Method on the CRAY X-MP/2*. Technical Report LBL-18023, University of California: Berkeley, 1984.
- Meurant G. *Computer Solution of Large Linear Systems*. North Holland: Amsterdam, 1999.
- Moler CB and Stewart GW. An algorithm for generalized matrix eigenvalue problems. *SIAM J. Numer. Anal.* 1973; 10:241–256.
- Munksgaard N. Solving sparse symmetric sets of linear equations by preconditioned conjugate gradient method. *ACM Trans. Math. Softw.* 1980; 6:206–219.
- Paige CC and Saunders MA. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.* 1975; 12:617–629.
- Parlett BN. *The Symmetric Eigenvalue Problem*. Prentice Hall: Englewood Cliffs, 1980.
- Saad Y. A flexible inner-outer preconditioned GMRES algorithm. *SIAM J. Sci. Comput.* 1993; 14:461–469.
- Saad Y. ILUT: A dual threshold incomplete LU factorization. *Numer. Lin. Alg. Appl.* 1994; 1:387–402.
- Saad Y. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company: Boston, 1996.
- Saad Y and Schultz MH. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.* 1986; 7:856–869.
- Saad Y and van der Vorst HA. Iterative solution of linear systems in the 20th century. *J. Comput. Appl. Math.* 2000; 123:1–23.
- Steijsen CLG and van der Vorst HA. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.* 1996; 17:401–425.
- Sonneveld P. CGS: a fast Lanczos-type solver for nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.* 1989; 10:36–52.
- Stewart GW. Perturbation theory for the definite generalized eigenvalue problem. In *Recent Advances in Numerical Analysis*, de Boor C, Golub GH (eds). Academic Press: New York, 1978; 193–206.
- Stewart GW. *Matrix Algorithms, Vol. I: Basic Decompositions*. SIAM: Philadelphia, 1998.
- Stewart GW. *Matrix Algorithms, Vol. II: Eigensystems*. SIAM: Philadelphia, 2001.
- van der Vorst HA. Iterative solution methods for certain sparse linear systems with a non-symmetric matrix arising from PDE-problems. *J. Comput. Phys.* 1981; 44:1–19.
- van der Vorst HA. Large tridiagonal and block tridiagonal linear systems on vector and parallel computers. *Parallel Comput.* 1987; 5:45–54.
- van der Vorst HA. High performance preconditioning. *SIAM J. Sci. Statist. Comput.* 1989; 10:1174–1185.
- van der Vorst HA. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems. *SIAM J. Sci. Statist. Comput.* 1992; 13:631–644.
- van der Vorst HA. Computational methods for large eigenvalue problems. In *Handbook of Numerical Analysis*, vol. VIII, Claret PG, Lions JL, (eds). North Holland: Amsterdam, 2002; 3–179.
- van der Vorst HA. *Iterative Krylov Methods for Large Linear Systems*. Cambridge University Press: Cambridge, 2003.
- van der Vorst HA and Vuk C. GMRESR: A family of nested GMRES methods. *Numer. Lin. Alg. Appl.* 1994; 1:369–386.
- van Gijzen MB. *Iterative Solution Methods for Linear Equations in Finite Element Computations*. PhD thesis, Delft University of Technology, Delft, 1994.
- Watkins DS. *Fundamentals of Matrix Computations*. John Wiley & Sons: New York, 1991.
- Watkins DS. Some perspectives on the eigenvalue problem. *SIAM Rev.* 1993; 35:430–471.
- Watkins DS. QR-like algorithms for eigenvalue problems. *J. Comput. Appl. Math.* 2000; 123:67–83.
- Wilkinson JH. *The Algebraic Eigenvalue Problem*. Clarendon Press: Oxford, 1965.

Chapter 20

Multigrid Methods for FEM and BEM Applications

Wolfgang Hackbusch

Max-Planck-Institut für Mathematik in den Naturwissenschaften, Inselstr., Leipzig, Germany

| | |
|---|-----|
| 1 General Remarks on Multigrid Methods | 577 |
| 2 Two-Grid Iteration | 581 |
| 3 Multigrid Method | 584 |
| 4 Application to Finite Element Equations | 586 |
| 5 Additive Variant | 589 |
| 6 Nested Iteration | 590 |
| 7 Nonlinear Equations | 592 |
| 8 Eigenvalue Problems | 593 |
| 9 Applications to the Boundary Element Method (BEM) | 593 |
| References | 595 |
| Further Reading | 595 |

1 GENERAL REMARKS ON MULTIGRID METHODS

1.1 Introduction

The solution of large systems of linear or even nonlinear equations is a basic problem when partial differential equations are discretized. Examples are the finite element equations in continuum or fluid dynamic problems. Since the dimension of these systems is often only limited by the available computer storage, the size of the arising systems is increasing because of the advances in computer technology. At the moment, systems with several millions of equations are of interest. The solution of such systems

requires numerical methods that have a runtime proportional to the dimension of the system (so-called 'linear complexity' or 'optimal complexity'). Immediately when computers started to be available, one tried to find more efficient solvers. Multigrid methods happened to be the first ones that reached the linear complexity for a rather large class of problems.

Because of its naming, multigrid methods involve several 'grids'. The nature of this grid hierarchy is explained below in a general setting. A simple one-dimensional example can be found in Section 1.2.5.

1.1.1 The standard problem structure

The situation we consider in the following is illustrated in diagram (1):

$$\begin{array}{c}
 \mathcal{P} = \mathcal{P}_{\text{continuous}} \\
 \downarrow \text{discretization process} \\
 \mathcal{P}_{\text{discrete}} = \mathcal{P}_\ell \longleftrightarrow \mathcal{P}_{\ell-1} \longleftrightarrow \dots \longleftrightarrow \mathcal{P}_0
 \end{array} \quad (1)$$

$\mathcal{P}_{\text{discrete}}$ is a given (discrete i.e. finitely dimensional) algebraic problem. The most prominent example of such a problem is a *system of linear equations*. Therefore, the discussion of the corresponding linear multigrid methods will fill the major part of this contribution. However, from the practical point of view, *nonlinear systems* may be a more important example of $\mathcal{P}_{\text{discrete}}$. Other examples are *eigenvalue problems* $\mathcal{P}_{\text{discrete}}$. It is essential for the multigrid approach to embed the discrete problem $\mathcal{P}_{\text{discrete}}$ into a *hierarchy* of discrete problems $\mathcal{P}_\ell, \mathcal{P}_{\ell-1}, \dots, \mathcal{P}_0$, where the problem size (dimension) of \mathcal{P}_k is increasing with increasing level number k . The largest dimension corresponds to $\mathcal{P}_{\text{discrete}} = \mathcal{P}_\ell$, while the lower dimensional problems

$\mathcal{P}_{\ell-1}, \dots, \mathcal{P}_0$ are used as auxiliary problems. The role of the continuous problem $\mathcal{P} = \mathcal{P}_{\text{continuous}}$ will be discussed below.

1.1.2 What are multigrid methods?

The multigrid method can be considered as a concept for the construction of fast iterative methods based on a hierarchy $\mathcal{P}_\ell, \mathcal{P}_{\ell-1}, \dots, \mathcal{P}_0$ for solving an algebraic problem $\mathcal{P}_{\text{discrete}} = \mathcal{P}_\ell$. To have an analogue, we may look at the finite element method (FEM). The FEM is not a special discretization of a particular problem but offers a whole class of discretizations (elements of different shape and order) to various problems (differential as well as integral equations). Similarly, the multigrid technique is applicable to large classes of (discrete algebraic) problems and describes how to construct algorithms for solving these problems. Instead of 'multigrid method' other names are also in use, for example, 'multiple grid method'. The name states that several 'grids' are involved, which belong to different levels of a hierarchy, as indicated in (1). This also explains the alternative namings 'multilevel method' or 'multiple level method'. The word 'grid' in 'multigrid method' originates from finite difference approximations by means of a regular grid. However, this does not mean that the multigrid method is restricted to such discretizations. In fact, the major part of this contribution will refer to the finite element discretization. Finite differences, finite elements or finite volumes are examples for the discretization process in (1).

1.1.3 Which problems can be solved?

The continuous problem $\mathcal{P} = \mathcal{P}_{\text{continuous}}$ can be connected with a partial differential or integral equation of elliptic type.

Usually, the problems $\mathcal{P}_{\text{discrete}}$ to be solved by multigrid are discrete analogues of a continuous problem $\mathcal{P} = \mathcal{P}_{\text{continuous}}$ derived by some discretization indicated by \downarrow in (1).

A given discretization process offers an easy possibility to obtain not only one particular discrete problem \mathcal{P} but a whole hierarchy of many discrete analogues \mathcal{P}_k ($k = \ell, \ell-1, \dots, 0$) of the continuous problem corresponding to different dimensions. This hierarchy is needed in the multigrid iteration.

The essence of the multigrid method is the so-called 'coarse grid correction'. This requires the existence of a lower dimensional problem $\mathcal{P}_{\ell-1}$, which approximates the given problem \mathcal{P}_ℓ in a reasonable way. If the discretization process is not given in advance, it must be created as a part of the multigrid process.

On the other hand, it is often very hard to apply multigrid, when the desired solution cannot be represented by a lower

dimensional approximation. Examples are boundary value problems with a geometrically complicated boundary (see end of Section 1.1.6)

1.1.4 Why is multigrid optimal?

If a multigrid method is successful, the iterative process has the following characteristics:

- (i) The convergence speed is uniform with respect to the problem size. Therefore, differently from simple iterative methods, which become slower with increasing dimension, these algorithms can be used for large scale problems. The desired accuracy can be obtained by a fixed number of multigrid steps. The resulting method is of linear complexity.
- (ii) Owing to the hopefully good approximation of problem \mathcal{P}_ℓ by $\mathcal{P}_{\ell-1}$, the convergence speed is expected to be much better (i.e. smaller) than 1. This implies that only very few iteration steps are necessary to obtain the desired accuracy.

The characteristics given above describe the aim we want to obtain with a good multigrid implementation. In fact, optimal convergence can be guaranteed in rather general cases. Here, 'general' means that no special algebraic properties of the matrix are required. In particular, the matrix may be neither symmetric nor positive definite.

Nevertheless, there is no easy way to get such results in any case. For singularly perturbed problems (e.g. problems with high Reynolds numbers) the convergence speed $\rho = \rho(\delta)$ may depend on a parameter δ . If $\rho(\delta)$ approaches 1 in the limit $\delta \rightarrow 0$ (or $\delta \rightarrow \infty$), the fast convergence promised in (ii) is lost. These difficulties give rise to the investigation of the so-called *robust multigrid methods*, which by definition lead to uniform convergence rates $\rho(\delta)$.

1.1.5 Is multigrid easy to implement?

As seen above, multigrid methods require an environment of a hierarchy consisting of the problems \mathcal{P}_k ($k = \ell, \ell-1, \dots, 0$) together with interacting mappings between neighboring levels $k, k-1$ (these interactions are denoted by \rightleftharpoons in (1)). If this environment is present in the implementation anyway (e.g. as a part of the adaptive refinement process), the multigrid method is rather easy to implement. If, however, only the problem description of $\mathcal{P}_{\text{discrete}}$ is given, the auxiliary problems \mathcal{P}_k for $k < \ell$ together with the interactions \rightleftharpoons must be installed as a part of the multigrid method, which of course makes the multigrid implementation much more involved.

1.1.6 The hierarchy of problems

There are different possibilities to produce a hierarchy of discretizations. For discretization methods based on regular grids, the hierarchy of grids induces the necessary hierarchy of discrete problems \mathcal{P}_k . In the case of FEMs, the underlying triangulation replaces the grid. A hierarchy of nested triangulations yields a perfect problem hierarchy. Such a hierarchy may be the side-product of adaptive mesh refinement. In this case, one proceeds from the coarsest to the finest level. The problem \mathcal{P}_0 corresponding to the coarsest mesh size should have a dimension small enough to be solved by standard methods. This can cause severe difficulties for complicated boundary value problems that need a high number of degrees of freedom (see e.g. Hackbusch and Sauter, 1997 and Section 3.4).

1.1.7 Notations

The norm $\|\cdot\|$ is the Euclidean norm when applied to vectors and the spectral norm for matrices (i.e. $\|A\| := \max \{\|Ax\| / \|x\| : x \neq 0\}$). $\langle \cdot, \cdot \rangle_k$ denotes the Euclidean scalar product of the vector space \mathcal{U}_k (see below). The scalar product in $L^2(\Omega)$ is denoted by $(f, g)_{L^2(\Omega)} := \int_\Omega f g \, dx$.

The Landau symbol $\mathcal{O}(f(x))$ means that the quantity is bounded by $C * f(x)$ for some constant C , when x tends to its characteristic limit (e.g. $x = h \rightarrow 0$ for the step size h or $x = n \rightarrow \infty$ for the dimension n). Further notations are explained below.

1.1.8 Literature

The first two-grid method is described by Brakhage (1960) for the solution of an integral equation (see Section 9.1). The first two-grid method for the Poisson equation is introduced by Fedorenko (1961), while Fedorenko (1964) contains the first multi-grid method. The first more general convergence analysis is given by Bakhvalov (1966).

Since there is a vast literature about multigrid methods, we do not try to give a selection of papers. Instead we refer to the monographs Hackbusch (1985), Wesseling (1991) (in particular devoted to problems of fluid dynamics), Bramble and Zhang (1993), Trottenberg, Oosterlee and Schüller (2001) (and the literature cited therein), and to the proceedings of the *European Multigrid Conferences* edited by Hackbusch and Trottenberg (1982) (containing an introduction to multigrid), Hackbusch and Trottenberg (1986), Hackbusch and Trottenberg (1991), Hemker and Wesseling (1994), and Hackbusch and Wittum (1998).

| Item | Explanation | Reference |
|---|--|-------------------------------|
| d_k | Defect $L_k u_k - f_k$ | (15) |
| f_k | vector of right-hand side (at level k) | (3) |
| h_k | grid size, mesh size | Section 1.2.1 |
| $\mathcal{H}_k, \mathcal{H}_k^{\text{FEM}} \subset \mathcal{H}$ | finite element space \subset energy space | Section 4, (33) |
| k, ℓ | level index | Section 1.2.1 |
| L_k | stiffness matrix of level k | (3), Section 4.1, Section 4.2 |
| $M_k, M_k^{\text{DOM}}, \dots$ | iteration matrix | Section 2.3 |
| p | prolongation | (5) |
| P, P_k | bijection onto FE space (of level k) | Section 4.1, Section 4.2 |
| r | restriction | (6) |
| S_k | iteration matrix corresponding to S_k | Section 2.1 |
| S_k, S_k^v | smoothing iteration, v -fold application of S_k | Section 1.2.2, Section 2.2 |
| T, T_k | triangulation (at level k) | Section 4.1, Section 4.2 |
| u, u_k | solution vector (at level k) | (3), Section 4.2 |
| u^k | finite element function at level k from \mathcal{H}_k | Section 4.3 |
| $\mathcal{U}, \mathcal{U}_k$ | linear space of vectors u, u_k | (5), Section 4.1 |
| γ | $\gamma = 1$: V-cycle, $\gamma = 2$: W-cycle | (27) |
| $\delta_{\alpha\beta}$ | $\delta_{\alpha\beta} = 0$ for $\alpha \neq \beta$, $\delta_{\alpha\alpha} = 1$ otherwise | Kronecker symbol |
| ζ, ζ_k | contraction number | Section 2.3, (50) |
| $\rho(\cdot)$ | spectral radius of a matrix | Section 2.3 |
| Ω | domain of boundary value problem | Section 4 |
| Ω_k | grid of level k , set of nodal values | Section 1.2.5 |

1.2 Ingredients of multigrid iterations

1.2.1 Hierarchy of linear systems

We consider the solution of a linear system

$$\mathbf{L}_k \mathbf{u}_k = \mathbf{f}_k \quad (2)$$

where h refers to the smallest mesh size $h = h_\ell$ of a hierarchy of mesh sizes $h_0 > h_1 > \dots > h_{\ell-1} > h_\ell$. We write the linear system corresponding to the level k (mesh size h_k) as

$$\mathbf{L}_k \mathbf{u}_k = \mathbf{f}_k \quad \text{for } k = \ell, \ell-1, \dots, 0 \quad (3)$$

In particular, $\mathbf{L}_\ell \mathbf{u}_\ell = \mathbf{f}_\ell$ is an identical writing for (2).

1.2.2 Smoothing iteration

Classical iterative methods like the Jacobi or the Gauss-Seidel iteration are needed for all levels $k > 0$. The purpose is not to reduce the iteration error but to produce smoother errors. This is the reason for the name 'smoothing iteration' or 'smoother'. The smoother is applied to the old iterate $\mathbf{u}_k^{\text{old}}$ and the right-hand side \mathbf{f}_k and produces the new iterate $\mathbf{u}_k^{\text{new}} = \mathbf{S}_k(\mathbf{u}_k^{\text{old}}, \mathbf{f}_k)$.

To give an example, we mention the simplest smoother, the Richardson iteration, which may be chosen in several cases,

$$\mathbf{S}_k(\mathbf{u}_k, \mathbf{f}_k) \mapsto \mathbf{u}_k - \frac{1}{\|\mathbf{L}_k\|}(\mathbf{L}_k \mathbf{u}_k - \mathbf{f}_k) \quad (4)$$

1.2.3 Prolongations

We denote the space of vectors \mathbf{u}_k and \mathbf{f}_k by \mathcal{U}_k . The prolongation $p_{k,k-1}$ is a linear transfer from \mathcal{U}_{k-1} to \mathcal{U}_k ,

$$p_{k,k-1}: \mathcal{U}_{k-1} \rightarrow \mathcal{U}_k \quad \text{linear} \quad (5)$$

In the following, we omit the indices $k, k-1$ and write $p: \mathcal{U}_{k-1} \rightarrow \mathcal{U}_k$.

1.2.4 Restrictions

The restriction $r_{k-1,k}$ is a linear transfer in the opposite direction,

$$r_{k-1,k}: \mathcal{U}_k \rightarrow \mathcal{U}_{k-1} \quad \text{linear} \quad (6)$$

Again, we abbreviate $r_{k-1,k}$ by r .

1.2.5 A one-dimensional example

Consider the 1D Dirichlet boundary value problem

$$\begin{aligned} -u''(x) &= f(x) \quad \text{in } \Omega = (0, 1) \\ u(x) &= 0 \quad \text{at } x \in \Gamma = \{0, 1\} \end{aligned} \quad (7)$$

Given a mesh size $h = 1/(n+1)$ with n being a power of 2, we define the grid Ω_h consisting of the points $x_v = vh$ ($v = 1, \dots, n$). The hierarchy of step sizes is

$$h_0 > h_1 > h_2 > \dots > h_\ell = h \quad \text{with } h_k = 2^{-k} \quad (8)$$

Note that $h_0 = 1/2$ is the largest possible step size. The number of grid points in $\Omega_k = \{x_v = vh_k: 1 \leq v \leq n_k\}$ is $n_k = 2^{k+1} - 1$ (see Figure 1).

The standard difference method yields the equations $h_k^{-2}[-u_{v-1} + 2u_v - u_{v+1}] = f(vh_k)$ for $v = 1, \dots, n_k$ with boundary values $u_0 = u_{n_k+1} = 0$. The matrices \mathbf{L}_k from (3) are

$$\mathbf{L}_k = h_k^{-2} \text{tridiag}\{-1, 2, -1\} \quad (k = 0, \dots, \ell) \quad (9)$$

of size $n_k \times n_k$, while the vectors are $\mathbf{u}_k = (u_v)_{v=1}^{n_k}$, $\mathbf{f}_k = (f(vh_k))_{v=1}^{n_k}$.

The prolongation $p = p_{k,k-1}$ from (5) can be realized by piecewise linear interpolation:

$$(p v_{k-1})(x) := \begin{cases} v_{k-1}(x) & \text{if } x \in \Omega_{k-1} \subset \Omega_k \\ \frac{v_{k-1}(x-h_k) + v_{k-1}(x+h_k)}{2} & \text{if } x \in \Omega_k \setminus \Omega_{k-1} \end{cases} \quad (10)$$

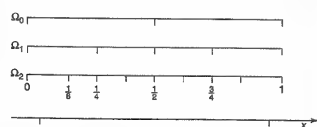


Figure 1. Grid hierarchy.

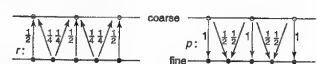


Figure 2. Prolongation and restriction in the 1D case.

(see Figure 2). p is described by the left rectangular matrix of size $n_k \times n_{k-1}$ in

$$p = \frac{1}{2} \begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ & 2 & 2 & \\ & & \ddots & \ddots \\ & & & 1 & 2 & 1 \\ & & & & 1 & 2 & 1 \end{bmatrix} \quad (11)$$

The restriction $r = r_{k,k-1}$ from (6) is defined by the following weighted mean value

$$(r v_{k-1})(x) = \frac{1}{4}[v_k(x-h_k) + 2v_k(x) + v_k(x+h_k)] \quad \text{for } x \in \Omega_{k-1} \quad (12)$$

(see Figure 2). The corresponding rectangular matrix of size $n_{k-1} \times n_k$ is the second in (11).

A suitable smoother is Jacobi's iteration damped by the factor $(1/2)$: $\mathbf{u}_k^{j+1} = \mathbf{S}_k(\mathbf{u}_k^j, \mathbf{f}_k)$ with

$$\mathbf{S}_k(\mathbf{u}_k, \mathbf{f}_k) := \mathbf{u}_k - \frac{1}{2} \mathbf{D}_k^{-1}(\mathbf{L}_k \mathbf{u}_k - \mathbf{f}_k) \quad (13)$$

where \mathbf{D}_k is the diagonal of \mathbf{L}_k , that is, $\mathbf{D}_k = 2h_k^{-2} \mathbf{I}$ (cf. (9)). Note that \mathbf{S}_k from (13) is almost identical to (4), since $(1/\|\mathbf{L}_k\|) \approx (1/4)h_k^2$.

2 TWO-GRID ITERATION

The two-grid iteration is a preversion of the multigrid iteration. It is not of practical interest, but it is an important building block that involves only the fine grid and one coarse grid.

In the following, we explain the so-called smoothing effect of certain classical iterations. This gives rise to the smoothing iteration mentioned in Section 1.2.2.

2.1 Smoothing effect

Usually, the purpose of an iteration $\mathbf{u}_k^0 \mapsto \mathbf{u}_k^1 \mapsto \dots \mapsto \mathbf{u}_k^j$ is the fast convergence to the true solution, that is, that the error $\mathbf{u}_k^j - \mathbf{u}_k$ decreases quickly. The purpose of a smoothing iteration is different. The error $\mathbf{u}_k^j - \mathbf{u}_k$ may stay as large as the starting error $\mathbf{u}_k^0 - \mathbf{u}_k$, but it must become smoother (note that the error $\mathbf{u}_k^j - \mathbf{u}_k$ must become smooth;

this does not concern the smoothness of the solution \mathbf{u}_k). The details are exemplified for the 1D example from Section 1.2.5.

We consider the damped Jacobi iteration (13) since its analysis is the most transparent one. Since $\mathbf{D}_k^{-1} = (1/2)h_k^2 \mathbf{I}$, the damped Jacobi iteration equals

$$\mathbf{u}_k^{j+1} = \mathbf{u}_k^j - \omega h_k^2 (\mathbf{L}_k \mathbf{u}_k^j - \mathbf{f}_k) = \mathbf{S}_\omega \mathbf{u}_k^j + \omega h_k^2 \mathbf{f}_k \quad (14)$$

$$\text{with } \omega = \frac{1}{4}$$

where $\mathbf{S}_\omega = \mathbf{I} - \omega h_k^2 \mathbf{L}_k$ is called the iteration matrix. Note that the eigenvectors of \mathbf{L}_k are

$$\mathbf{e}_\mu^k = \sqrt{2h_k} (\sin(\mu \pi h_k))_{v=1}^{n_k} \quad \text{for } \mu = 1, \dots, n_k$$

(μ : frequency). The corresponding eigenvalues of \mathbf{L}_k are $\lambda_{k,\mu} = 4h_k^{-2} \sin^2(\mu \pi h_k/2)$, that is, $\mathbf{L}_k \mathbf{e}_\mu^k = \lambda_{k,\mu} \mathbf{e}_\mu^k$. Since $\mathbf{S}_\omega = \mathbf{I} - \omega h_k^2 \mathbf{L}_k$ is the iteration matrix of the damped iteration (14), it has the same eigenvectors \mathbf{e}_μ^k as \mathbf{L}_k and (for $\omega = (1/4)$) the eigenvalues

$$\begin{aligned} \lambda_\mu &= 1 - \sin^2\left(\frac{\mu \pi h_k}{2}\right) = \cos^2\left(\frac{\mu \pi h_k}{2}\right) \\ (1 \leq \mu \leq n_k = h_k^{-1} - 1) \end{aligned}$$

shown in Figure 3. The choice $\omega = (1/2)$ yields the standard Jacobi (dashed line in Figure 3).

The rate of convergence of the damped Jacobi iteration is $\lambda_1 = \cos^2(\pi h_k/2) = 1 - (1/4)\pi^2 h_k^2 + \mathcal{O}(h_k^4)$, proving the very slow convergence of the Jacobi iteration.

Even though iteration (14) converges very slowly, Figure 3 shows that components \mathbf{e}_μ^k with frequency $\mu \geq 1/(2h_k)$ are reduced at least by a factor $(1/2)$ per iteration. This means that the convergence rate of the damped Jacobi iteration restricted to the subspace span $\{\mathbf{e}_\mu^k: 1/2 \leq \mu h_k < 1\}$ of the high frequencies is $1/2$. The iteration is rapidly convergent with respect to the high frequencies. The slow convergence is caused by the lower frequencies

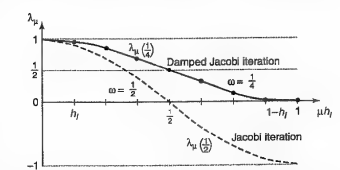


Figure 3. Eigenvalues of the iteration matrix as function of frequency μh_k .

only. The construction of the multigrid iteration is based on this observation.

The initial error $u_0^i - u_k$ can be represented by the linear combination $\sum \alpha_k e_k^i$. After v steps of the damped Jacobi iteration, the error equals $u_1^i - u_k = \sum \beta_k e_k^i$ with $\beta_k = \alpha_k \lambda_k^v$. The preceding consideration shows $\beta_k \approx \alpha_k$ for low frequencies but $|\beta_k| \ll |\alpha_k|$ for high frequencies. This fact can be expressed by saying that the error $u_1^i - u_k$ is smoother than $u_0^i - u_k$. The first three graphs of Figure 4 illustrate the increasing smoothness of $u_k^i - u_k$ ($v = 0, 1, 2$). This is why we say that the iteration (14) serves as a smoothing iteration.

2.2 Structure of two-grid iterations

The foregoing subsection showed that an appropriate smoothing iteration is quite an efficient method for reducing the high-frequency components. Convergence is only lacking with respect to low frequencies (smooth components). Therefore, one should combine this iteration with a second one having complementary properties. In particular, the second iteration should reduce the smooth error part very well. Such a complementary iteration can be constructed by means of the coarse grid with step size $h_{\ell-1} = 2h_\ell$ (see Figure 1).

Let u_k^{old} be some given approximation to $u_k = L_\ell^{-1} f_\ell$, which serves as starting value. Few steps of the smoothing iteration S_ℓ (cf. (13)) will result in an intermediate value \bar{u}_ℓ . From the previous subsection, we know that the error $v_\ell = \bar{u}_\ell - u_k$ is smooth (more precisely, smoother than $u_k^{\text{old}} - u_k$). v_ℓ can also be regarded as the exact correction since $u_k = \bar{u}_\ell - v_\ell$ represents the exact solution. Inserting \bar{u}_ℓ into the equation $L_\ell u_k - f_\ell = 0$, we obtain the defect

$$d_\ell = L_\ell \bar{u}_\ell - f_\ell \quad (15)$$

of \bar{u}_ℓ , which vanishes if and only if \bar{u}_ℓ is the exact solution u_k . Because of $L_\ell v_\ell = L_\ell \bar{u}_\ell - L_\ell u_k = L_\ell \bar{u}_\ell - f_\ell = d_\ell$,

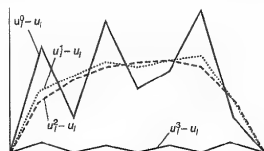


Figure 4. Smoothing effect of the damped Jacobi iteration. u_k : exact discrete solution; u_k^i , u_k^j , u_k^l : smoothing iterates; u_k^i : result after coarse grid correction.

the exact correction v_ℓ is the solution of

$$L_\ell v_\ell = d_\ell \quad (16)$$

Equation (16) is of the same form as the original equation $L_\ell u_k = f_\ell$. Solving $L_\ell v_\ell = d_\ell$ exactly is as difficult as solving $L_\ell u_k = f_\ell$. Nonetheless, v_ℓ can be approximated better than u_k , since v_ℓ is smooth and smooth functions can be represented well by means of coarser grids.

To approximate the problem $L_\ell v_\ell = d_\ell$ by a coarse grid equation

$$L_{\ell-1} v_{\ell-1} = d_{\ell-1} \quad (17)$$

we have to choose $d_{\ell-1}$ reasonably. Note that the matrix L_ℓ is already defined by (9) for all levels $k \geq 0$, especially for $k = \ell - 1$. The right-hand side $d_{\ell-1}$ should depend linearly on the original defect d_ℓ . This is where the restriction r from (6) is needed,

$$d_{\ell-1} := r d_\ell \quad (18)$$

Having defined $d_{\ell-1}$, we obtain $v_{\ell-1} = L_{\ell-1}^{-1} d_{\ell-1}$ as the exact solution of (17). We expect $v_{\ell-1}$ to be an approximation of the exact correction v_ℓ . However, $v_{\ell-1}$ is only defined on the coarse grid $\Omega_{\ell-1}$. We have to interpolate this coarse grid function by

$$\bar{v}_\ell = p v_{\ell-1} \quad (19)$$

where the prolongation p is announced in (5) (see Figure 2).

Since $u_k = \bar{u}_\ell - v_\ell$ is the exact solution and $\bar{v}_\ell = p v_{\ell-1}$ is supposed to approximate v_ℓ , one tries to improve the value \bar{u}_ℓ by

$$u_k^{\text{new}} := \bar{u}_\ell - \bar{v}_\ell \quad (20)$$

The step from u_k to u_k^{new} by (15)–(20) is called the coarse grid correction. Combining the separate parts (15)–(20), we obtain the compact formula

$$\bar{u}_\ell \mapsto \bar{u}_\ell - p L_{\ell-1}^{-1} r (L_\ell \bar{u}_\ell - f_\ell) \quad (21)$$

for the coarse grid correction.

Figure 4 shows the errors $u_k^i - u_k$ after $i = 0, 1, 2$ damped Jacobi iterations. The coarse grid correction applied to $\bar{u}_\ell = u_k^2$ yields u_k^3 . The graph of the error $u_k^3 - u_k$ in Figure 4 proves the success of the coarse grid correction. Although the coarse grid correction seems to be efficient, it cannot be used as an iteration by itself because it does not converge. It is the combination of smoothing iteration and coarse grid correction that is

Two-grid iteration for solving $L_\ell u_\ell = f_\ell$

Start: u_ℓ^j given iterate

Smoothing step:

$$\bar{u}_\ell := S_\ell^v(u_\ell^j, f_\ell) \quad (v \text{ smoothing steps}) \quad (a)$$

Coarse grid correction:

$$d_\ell := L_\ell \bar{u}_\ell - f_\ell \quad (\text{calculation of the defect}) \quad (b_1)$$

$$d_{\ell-1} := r d_\ell \quad (\text{restriction of the defect}) \quad (b_2)$$

$$v_{\ell-1} := L_{\ell-1}^{-1} d_{\ell-1} \quad (\text{solution of coarse grid eq.}) \quad (b_3)$$

$$u_\ell^{j+1} := \bar{u}_\ell - p v_{\ell-1} \quad (\text{correction of } \bar{u}_\ell) \quad (b_4)$$

(22)

rapidly convergent, whereas both components by themselves converge slowly or not at all. The combination is called the two-grid iteration since two levels ℓ and $\ell - 1$ are involved.

We summarize the two-grid iteration in (22). In Step (a), we use the notation $S_\ell^v(u_\ell^j, f_\ell)$ for the v -fold application of the smoothing iteration $S_\ell(u_\ell^j, f_\ell)$.

The number v of smoothing iterations can be chosen independently of the grid size h_ℓ . Its influence on convergence will be described in Remark 1.

Below, the two-grid iteration (22) is written in a quasi-ALGOL style.

The function TGM performs one iteration step at level k (first parameter). The third parameter f is the right-hand side f_k of the equation to be solved. The input value of the second parameter u is the given j th iterate u_k^j that is mapped into the output value $TGM = u_k^{j+1}$. The second line 'if $\ell = 0$ then ...' is added to have a well-defined algorithm for all levels $k \geq 0$. Note that $TGM = TGM^{(v)}$ depends on the choice of v .

Iteration (22) can be regarded as the prototype of a two-grid method. However, there are many variants. Instead of applying first smoothing and thereafter the coarse grid correction, we can interchange these parts. More generally, v_1 smoothing iterations can be performed before and v_2 iterations after the coarse grid correction. The resulting algorithm is given below in (23).

2.3 Two-grid convergence in the 1D case

In the case of the one-dimensional model problem from Section 1.2.5, a discrete Fourier analysis can be applied. The explicit calculation can be found in Hackbusch (1985) and in Section 10.3 Hackbusch (1994). Here we give only the results.

The iteration matrix $M_\ell = M_\ell^{TGM(v)}$ of the two-grid method (22) is defined by $u_\ell^{j+1} = M_\ell u_\ell^j + N_\ell f_\ell$ and equals

$$M_\ell^{TGM(v)} = (I - p L_{\ell-1}^{-1} r L_\ell) S_\ell^v$$

```
function TGM(k, u, f);
begin if k = 0 then TGM := L_0^{-1} * f else
  begin u := S_k^v(u, f);
    d := r * (L_k * u - f);
    v := L_{k-1}^{-1} * d;
    TGM := u - p * v
  end end;
```

(22')

```
function TGM(k, u, f);
begin if k = 0 then TGM := L_0^{-1} * f else
  begin u := S_k^v(u, f);
    u := u - p * L_{k-1}^{-1} * r * (L_k * u - f);
    TGM := S_k^v(u, f)
  end end;
```

(23)

Figure 6. One V-cycle ($\gamma = 1$; left) and one W-cycle ($\gamma = 2$; right) for $\ell = 4$.

with Dirichlet values $u(x, y) = x^2 + y^2$. The coarsest possible mesh size is $h_0 := 1/2$, the further mesh sizes are $h_k = 2^{-k-1}$. The table shows the iteration errors $e_m := \|u_m^j - u_m^{j+1}\|_\infty$ after m steps of the multigrid iteration ($u_0^j = 0$, W-cycle, 2 Gauss-Seidel pre-smoothing steps) at level $\ell = 7$ (corresponding to $h_\ell = 1/256$ and 65 025 degrees of freedom).

| m | e_m | ratio |
|-----|--------------------|--------|
| 0 | $1.984_{10^{-9}}$ | — |
| 1 | $3.038_{10^{-1}}$ | 0.1531 |
| 2 | $1.605_{10^{-2}}$ | 0.0528 |
| 3 | $9.017_{10^{-4}}$ | 0.0562 |
| 4 | $5.219_{10^{-5}}$ | 0.0579 |
| 5 | $3.102_{10^{-6}}$ | 0.0594 |
| 6 | $1.884_{10^{-7}}$ | 0.0607 |
| 7 | $1.166_{10^{-8}}$ | 0.0619 |
| 8 | $7.713_{10^{-10}}$ | 0.0662 |
| 9 | $5.218_{10^{-11}}$ | 0.0677 |

The column 'ratio' shows the error reduction corresponding to a convergence rate of 0.067. Similar rates are observed for other step sizes. The starting value $u_0^j = 0$ is chosen here only for demonstration. Instead it is recommended to use the nested iteration from Section 6.

The results from above correspond to the W-cycle. The V-cycle shows a somewhat slower convergence rate of about $1.8_{10^{-1}}$ (the error for $m = 9$ is $e_9 = 4.98_{10^{-9}}$). On the other hand, the V-cycle requires less computational work (see Remark 3). The choice $\gamma = 3$ is quite impractical. Although the computational work increases a lot, $\gamma = 3$ yields almost the same results as the W-cycle, for example, $e_9 = 4.793_{10^{-11}}$.

3.3 Computational work

Let $n_k = \dim U_k$ be the number of degrees of freedom at level k . Assuming $h_{k-1} \approx 2h_k$ and $\Omega \subset \mathbb{R}^d$, we have $n_k \approx 2^d n_{k-1}$. Owing to recursivity, one call of $MGM(\ell, \cdot, \cdot)$ involves $\gamma^{\ell-k}$ calls of $MGM(k, \cdot, \cdot)$. As long as $\gamma 2^{-d} < 1$, the number of arithmetical operations spent at level k decreases exponentially with decreasing level. Indeed, even $\gamma 2^{-d} \leq 1/2$ holds because of $d \geq 2$ ($d = 1$ is uninteresting) and $\gamma \leq 2$ (V/W-cycle). Therefore, linear complexity holds.

Remark 3. Let $n_k \approx 2^d n_{k-1}$. The cost of one call of $MGM(\ell, \cdot, \cdot)$ is bounded by $C_\ell n_\ell$ arithmetical operations,

where for $d = 2$

$$C_\ell \leq \begin{cases} \frac{4}{3} [vC_S + C_D + C_C] + O(4^{-\ell}) & \text{for } \gamma = 1 \text{ (V-cycle),} \\ 2[vC_S + C_D + C_C] + O(2^{-\ell}) & \text{for } \gamma = 2 \text{ (W-cycle)} \end{cases} \quad (29)$$

where $C_S n_\ell$, $C_D n_\ell$, $C_C n_\ell$ and $C_0 n_0$ are the costs for $u_\ell \mapsto S_\ell(u_\ell, f_\ell)$, $(u_\ell, f_\ell) \mapsto r(L_\ell u_\ell - f_\ell)$, $(u_\ell, v_{\ell-1}) \mapsto u_\ell - p v_{\ell-1}$ and $f_0 \mapsto L_0^{-1} f_0$. For $d = 3$, the factors (4/3) and 2 in (29) reduce to (4/3) and (8/7), respectively.

3.4 Algebraic multigrid methods

So far we have assumed that there is hierarchy of discretizations (cf. Section 1.1.6). If only the final discretization is given or by some other reason the user is not providing coarser systems, this construction can be performed as part of the 'algebraic multigrid method' (AMG). There are different approaches how to select a coarse grid and how to define the prolongations and restrictions. The name 'algebraic multigrid method' indicates that only the algebraic system of equations is used as input (not necessarily geometric data and descriptions about the nature of the PDE). As a consequence, the resulting algorithm is more 'black-box'-like.

We refer the interested reader to Stüben (1983), Braess (1995), Mandel, Brezina and Vanek (1999), Haase et al. (2001), and the literature given therein.

4 APPLICATION TO FINITE ELEMENT EQUATIONS

While the introductory example corresponds to a difference scheme, we now discuss the multigrid method in the case of a finite element discretization. The multigrid ingredients from Section 1.2 are defined in a canonical way, provided that the finite element spaces are nested as explained below. Otherwise, hints are given in Section 4.5.

We assume that the boundary value problem is formulated in the weak variational form: Find $u \in \mathcal{H}$ such that

$$a(u, v) = f(v) \quad \text{for all } v \in \mathcal{H} \quad (30)$$

where the 'energy space' \mathcal{H} may include the required homogeneous Dirichlet conditions (e.g., $\mathcal{H} = H_0^1(\Omega)$). In the case of scalar functions, u the bilinear form $a(u, v)$

may be

$$a(u, v) := \int_\Omega \left(\sum_{\alpha, \beta=1}^d c_{\alpha, \beta} \frac{\partial u}{\partial x_\alpha} \frac{\partial v}{\partial x_\beta} + \sum_{\alpha=1}^d c_{\alpha, 0} \frac{\partial u}{\partial x_\alpha} v + \sum_{\beta=1}^d c_{0, \beta} u \frac{\partial v}{\partial x_\beta} + c_{0, 0} u v \right) dx$$

The functional f in (30) is $f(v) = \int_\Omega f v dx$. In the case of inhomogeneous Neumann conditions, it contains a further term $\int_\Gamma \psi v d\Gamma$ ($\Gamma = \partial\Omega$ denotes the boundary of Ω) (see Chapter 4, this Volume, Chapter 2, Volume 2).

4.1 Finite element problem

The FE space \mathcal{H}^{FEM} is a finite-dimensional subspace of \mathcal{H} and the FE problems reads:

$$\text{Find } u^{\text{FEM}} \in \mathcal{H}^{\text{FEM}} \quad \text{with} \quad a(u^{\text{FEM}}, v) = f(v) \quad \text{for all } v \in \mathcal{H}^{\text{FEM}} \quad (31)$$

The usual approach in the 2D case, is to define \mathcal{H}^{FEM} by means of piecewise linear (quadratic, ...) elements corresponding to a triangulation \mathcal{T} , which is a set of triangles such that $\bigcup_{\tau \in \mathcal{T}} \tau = \bar{\Omega}$. In 3D, the 'triangulation' \mathcal{T} contains tetrahedra and so on.

The functions $u \in \mathcal{H}^{\text{FEM}}$ are represented by means of the nodal basis, that is, there is a set \mathcal{I} of indices α associated with 'nodal points' $x_\alpha \in \mathbb{R}^d$ and a basis

$$\mathcal{B} = \{\phi_\alpha; \alpha \in \mathcal{I}\} \subset \mathcal{H}^{\text{FEM}}$$

with the interpolation property $\phi_\alpha(x_\beta) = \delta_{\alpha\beta}$ for all $\alpha, \beta \in \mathcal{I}$. Setting $u_\alpha := u(x_\alpha)$, we obtain

$$u = \sum_{\alpha \in \mathcal{I}} u_\alpha \phi_\alpha \quad \text{for any } u \in \mathcal{H}^{\text{FEM}} \quad (32)$$

The coefficients u_α form the coefficient vector $\mathbf{u} = (u_\alpha)_{\alpha \in \mathcal{I}} \in \mathcal{U}$. Relation (32) gives rise to the interpolation $\mathbf{u} \mapsto u := P\mathbf{u} \in \mathcal{H}^{\text{FEM}}$ by means of (32). P is a bijection from \mathcal{U} onto \mathcal{H}^{FEM} .

The stiffness matrix \mathbf{L} is given by the entries $L_{\alpha\beta} = a(\phi_\alpha, \phi_\beta)$, while the right-hand side vector is \mathbf{f} with $f_\alpha = f(\phi_\alpha)$. Altogether, we obtain the finite element equation $\mathbf{L}\mathbf{u} = \mathbf{f}$.

4.2 Nested finite element spaces

In order to get a family of finite element problems $L_k u_k = f_k$, we may assume a sequence of nested subspaces, that is,

there are finite element spaces \mathcal{H}_k ($0 \leq k \leq \ell$) replacing \mathcal{H}^{FEM} in (31) such that

$$\mathcal{H}_0 \subset \mathcal{H}_1 \subset \dots \subset \mathcal{H}_\ell \subset \mathcal{H} \quad (33)$$

This setting corresponds to conforming finite element methods. The nonconforming case will be considered in Section 4.6. But even in the conforming case (i.e. $\mathcal{H}_\ell \subset \mathcal{H}$ for all levels ℓ), one may use subspaces with $\mathcal{H}_{\ell-1} \not\subset \mathcal{H}_\ell$ (see Section 4.5).

The easiest way to construct these nested subspaces is by repeated refinement of the triangulation. Let \mathcal{T}_0 be the coarsest triangulation. The refined triangulations \mathcal{T}_k must satisfy the following:

each triangle $\tau \in \mathcal{T}_k$ is the union of triangles τ'_i , τ'_2, \dots from \mathcal{T}_{k+1} ($k = 0, \dots, \ell - 1$)

Then, elements being piecewise linear (quadratic) on $\tau \in \mathcal{T}_k$ are piecewise linear (quadratic) on $\tau' \in \mathcal{T}_{k+1}$; hence, $u \in \mathcal{H}_k$ belongs also to \mathcal{H}_{k+1} , that is, $\mathcal{H}_k \subset \mathcal{H}_{k+1}$ as required in (33).

We introduce the following notation: $u_k, f_k \in \mathcal{U}_k$ for vectors at level k , while \mathbf{L}_k is the stiffness matrix at level k . The nodal basis vectors are $\phi_{k,\alpha}$ ($\alpha \in \mathcal{I}_k$, $\mathcal{I}_k = \{0, \dots, \ell\}$), where \mathcal{I}_k is the index set of level k . The interpolation (32) is now denoted by $P_k: \mathcal{U}_k \rightarrow \mathcal{H}_k$.

4.3 Canonical prolongations and restrictions for FE equations

The multigrid prolongation $p: \mathcal{U}_{k-1} \rightarrow \mathcal{U}_k$ (see (5)) is uniquely defined as follows. Let $u_{k-1} \in \mathcal{U}_{k-1}$ a given coefficient vector and consider the corresponding finite element function $u^{k-1} = P_{k-1} u_{k-1} \in \mathcal{H}_{k-1}$. Since $\mathcal{H}_{k-1} \subset \mathcal{H}_k$, u^{k-1} allows a basis representation $u^{k-1} = P_k u_k$ (cf. (32)) by means of a unique coefficient vector $u_k \in \mathcal{U}_k$. Define $p u_{k-1}$ by u_k . Hence, the formal definition of p is $P_k^{-1} P_{k-1}$:

$$\begin{array}{ccc} \mathcal{U}_{k-1} & \xrightarrow{p} & \mathcal{U}_k \\ \downarrow P_{k-1} & & \downarrow P_k \\ \mathcal{H}_{k-1} & \xrightarrow{\text{inclusion}} & \mathcal{H}_k \end{array} \quad p = P_k^{-1} P_{k-1} \quad (34)$$

The canonical restriction is $r = p^*$, where p is the canonical prolongation from above, that is,

$$\langle r f_k, v_{k-1} \rangle_{k-1} = \langle f_k, p v_{k-1} \rangle_k \quad \text{for all } v_{k-1} \in \mathcal{U}_{k-1} \quad (35)$$

Here, $\langle \cdot, \cdot \rangle_k$ denotes the Euclidean scalar product of \mathcal{U}_k .

4.4 Coarse grid matrix and coarse grid correction

With a view to the next section, we state the characteristic relation between p , r , and the stiffness matrices L_k , L_{k-1} .

Remark 4. Let L_p and L_{k-1} be the finite element stiffness matrices of Section 4.1 and let p and r be canonical. Then L_{k-1} coincides with the Galerkin product

$$L_{k-1} = r L_k p \quad (36)$$

The coarse grid correction $C_k(u_k, f_k) := u_k - p L_{k-1}^{-1} r (L_k u_k - f_k)$ in (21) can be reformulated in terms of the finite element function $u^k = P_k u_k \in \mathcal{H}_k$ as follows.

Proposition 1. Assume $\mathcal{H}_{k-1} \subset \mathcal{H}_k$ and let p, r be the canonical mappings from Section 4.3. Let $a(\cdot, \cdot)$ be the bilinear form of the underlying variational formulation. Then, the coarse grid correction $u_k \mapsto u_k - p v_{k-1}$ from (27c₄) is equivalent to the mapping

$$\begin{aligned} u^k &\mapsto u^k - v^{k-1} \quad (37a) \\ a(v^{k-1}, w^{k-1}) &= a(u^k, w^{k-1}) - f(w^{k-1}) \end{aligned}$$

$$\text{for all } w^{k-1} \in \mathcal{H}_{k-1} \quad (37b)$$

where $v^{k-1} \in \mathcal{H}_{k-1}$ is the finite element solution of (37b). The solvability of (37b) is equivalent to the regularity of L_{k-1} .

The corresponding representation for nonconforming finite elements is given in (43).

4.5 Nonnested finite element spaces

So far, we have assumed the finite element spaces to be nested: $\mathcal{H}_0 \subset \mathcal{H}_1 \subset \dots \subset \mathcal{H}_\ell$. If \mathcal{H}_k are standard finite element spaces over a triangulation \mathcal{T}_k , the inclusion $\mathcal{H}_{k-1} \subset \mathcal{H}_k$ requires that all coarse triangles $i \in \mathcal{T}_{k-1}$ must be the union of some fine triangles $i_1, \dots, i_\ell \in \mathcal{T}_k$. This condition may be violated close to a curved boundary or near a curved interface. It may even happen that \mathcal{H}_{k-1} and \mathcal{H}_k are completely unrelated except that $\dim \mathcal{H}_{k-1} < \dim \mathcal{H}_k$ expresses that \mathcal{H}_{k-1} gives rise to a coarser discretization. In these cases, the definition (34) of the canonical prolongation cannot be applied.

The cheapest remedy is the definition of p by means of the finite element interpolation. Let $u_{k-1} \in \mathcal{U}_{k-1}$ be the coefficient vector corresponding to the nodal basis of \mathcal{H}_{k-1} , that is, $(P_{k-1} u_{k-1})(x_a) = u_{k-1,a}$ for $x_a \in \Omega_{k-1}$. Here, Ω_k is the set of nodal points of the triangulation \mathcal{T}_k . The desired

vector $p u_{k-1} \in \mathcal{U}_k$ has to represent the values $P_k p u_{k-1}$ on Ω_k . Even when $P_k p u_{k-1}$ cannot coincide with $P_{k-1} u_{k-1}$ in \mathcal{H}_k , it can interpolate $P_{k-1} u_{k-1}$:

$$(p u_{k-1})(x_a) := (P_{k-1} u_{k-1})(x_a) \quad \text{for all } x_a \in \Omega_k \quad (38)$$

The next remark ensures that definition (38) is a true generalization of the canonical choice.

Remark 5. If $\mathcal{H}_{k-1} \subset \mathcal{H}_k$, definition (38) yields the canonical prolongation p .

As soon as p is defined, we can define the restriction by (35), that means by $r = p^*$.

4.6 Nonconforming finite element discretizations

In the following, we give only the construction of the discretization and multigrid ingredients. For more details, we refer to Braess, Dryja and Hackbusch (1999).

4.6.1 Nonconforming discretization

Let $\mathcal{H}_k \subset L^2(\Omega)$ be a family of (nonconforming) finite element spaces, that is, we do not assume $\mathcal{H}_k \subset \mathcal{H}$. Moreover, the spaces \mathcal{H}_k are not supposed to be nested $\mathcal{H}_{k-1} \not\subset \mathcal{H}_k$. Instead of the bilinear form $a(\cdot, \cdot)$ a mesh-dependent bilinear form $a_k(\cdot, \cdot)$ on $\mathcal{H}_k \times \mathcal{H}_k$ is used. For $f \in \mathcal{H}^0$, the variational problem is discretized by

$$u^k \in \mathcal{H}_k \quad \text{with} \quad a_k(u^k, v^k) = f(v^k) \quad \text{for all } v^k \in \mathcal{H}_k \quad (39)$$

Again the isomorphism between \mathcal{U}_k and \mathcal{H}_k is denoted by P_k (cf. Section 4.3).

4.6.2 Multi-grid prolongation

The canonical prolongation $p = P_k^{-1} \circ P_{k-1}$ given by (34) is based on the inclusion $\mathcal{H}_{k-1} \subset \mathcal{H}_k$. Since $\mathcal{H}_{k-1} \not\subset \mathcal{H}_k$, the prolongation p must be defined differently. The inclusion $\mathcal{H}_{k-1} \subset \mathcal{H}_k$ is to be replaced by a suitable (linear) mapping

$$\iota: \mathcal{H}_{k-1} \rightarrow \mathcal{H}_k \quad (40)$$

Once ι is constructed, we are able to define the canonical prolongation and restriction by

$$p := P_k^{-1} \circ \iota \circ P_{k-1} \quad \text{and} \quad r := p^* = P_{k-1}^* \circ \iota^* \circ (P_k^*)^{-1} \quad (41)$$

Although the algorithm needs only the above mapping $\iota: \mathcal{H}_{k-1} \rightarrow \mathcal{H}_k$, it is easier to define ι on a larger space $\Sigma \supset \mathcal{H}_{k-1} + \mathcal{H}_k$ such that ι restricted to \mathcal{H}_{k-1} is the identity.

Since we only require $\Sigma \subset L^2(\Omega)$, no global smoothness is necessary.

Next, we need an auxiliary space S , connected with Σ and \mathcal{H}_k via the mappings σ and π , as shown in the following commutative diagram:

$$\begin{array}{ccccc} & & \Sigma & \xrightarrow{\sigma} & S \\ \text{inclusion } \uparrow & & \nwarrow & & \downarrow \pi \\ \mathcal{H}_{k-1} & \xrightarrow{\quad} & \mathcal{H}_k & & \\ P_{k-1} \uparrow & & \downarrow \iota & & \uparrow P_k \\ \mathcal{U}_{k-1} & \xrightarrow{\quad} & \mathcal{U}_k & & \end{array}$$

$\pi: S \rightarrow \mathcal{H}_k$ is required to be injective. The desired mapping ι (more precisely, its extension to Σ) is the product

$$\iota = \pi \circ \sigma: \Sigma \rightarrow \mathcal{H}_k \quad (42)$$

For the simplest nonconforming finite elements, the *Crouzeix-Raviart element*, we specify the spaces and mappings from above in

Example 1. Let \mathcal{T}_{k-1} be the coarse triangulation of the domain Ω , while \mathcal{T}_k is obtained by regular halving of all triangle sides. \mathcal{H}_k is the space of all piecewise linear functions that are continuous at the midpoints of edges in \mathcal{T}_k . Define the nodal point set $\Omega_k = \{x_\alpha; \alpha \in \mathcal{I}_k\}$ by all midpoints of edges in \mathcal{T}_k (except boundary points in the case of Dirichlet conditions). For all $x_\alpha \in \Omega_k$, basis functions $b_\alpha^k \in \mathcal{H}_k$ are defined by $b_\alpha^k(x_\beta) = \delta_{\alpha\beta}$ ($\alpha, \beta \in \mathcal{I}_k$). Then, \mathcal{U}_k is the coefficient space that is mapped by $P_k: \mathcal{U}_k = (\mathcal{U}_{k,\alpha})_{\alpha \in \mathcal{I}_k} \mapsto u^k = \sum_{\alpha \in \mathcal{I}_k} \mathcal{U}_{k,\alpha} b_\alpha^k$ onto \mathcal{H}_k . Similarly, \mathcal{H}_{k-1} , \mathcal{U}_{k-1} , and the isomorphism P_{k-1} are defined.

An appropriate space Σ , is the space of piecewise linear elements with respect to the fine triangulation \mathcal{T}_k that may be discontinuous. Obviously, $\mathcal{H}_{k-1} + \mathcal{H}_k \subset \Sigma$.

We set $S := \mathcal{U}_k$, $\pi := P_k$, and define σ as follows: Every nodal point $x_\alpha \in \Omega_k$ is the midpoint of the common side of adjacent triangles $i, i' \in \mathcal{T}_k$. We define the image σv by

$$(\sigma v)_\alpha := \frac{1}{2} [v|_i(x_\alpha) + v|_{i'}(x_\alpha)] \quad \text{for all } x_\alpha \in \Omega_k \text{ and } v \in \Sigma$$

Here, the linear function $v|_i$ is understood to be extended to the closure of i .

The multigrid prolongation is $p = \sigma \circ P_{k-1}$.

4.6.3 Coarse grid correction

Given p from (41) and $r := p^*$, the coarse grid correction takes the standard form (21). Its FE interpretation is as follows:

Let an FE approximation $u^k \in \mathcal{H}_k$ be given. Its defect $d^k \in \mathcal{H}_k$ is defined by

$$(d^k, w^k)_{L^2(\Omega)} = a_k(u^k, w^k) - f(w^k) \quad \text{for all } w^k \in \mathcal{H}_k$$

Using (39) and the error $e^k := u^k - u^*$, one obtains a characterization of d^k by

$$(d^k, w^k)_{L^2(\Omega)} = a_k(e^k, w^k) \quad \text{for all } w^k \in \mathcal{H}_k$$

Then the correction $e^{k-1} \in \mathcal{H}_{k-1}$ is determined as the solution of the FE coarse grid equation

$$a_{k-1}(e^{k-1}, w^{k-1}) = a_k(e^k, w^{k-1}) \quad \text{for all } w^{k-1} \in \mathcal{H}_{k-1} \quad (43)$$

Here ι is the mapping specified in (40). It is required for converting the function u^{k-1} from \mathcal{H}_{k-1} into a function in \mathcal{H}_k . The correction yields the new approximation $u^{k,\text{new}} := u^k - \iota e^{k-1}$.

5 ADDITIVE VARIANT

5.1 The additive multigrid algorithm

If one denotes the coarse grid correction (21) by $C_k(u_k, f_k) := u_k - p L_{k-1}^{-1} r (L_k u_k - f_k)$, the two-grid iteration is the product $C_k \circ S_k^*$, that is, $TGM(k, u_k, f_k) = C_k(S_k^*(u_k, f_k))$. Instead of the product, one can form the sum of the correction $\delta_{k,k} := u_k - S_k^*(u_k, f_k)$ of the smoothing procedure and the coarse grid correction $\delta_{k,k-1} := p L_{k-1}^{-1} r (L_k u_k - f_k)$. Damping the sum of these terms by a factor ϕ , one obtains the iteration $u_k^i \mapsto u_k^{i+1} := u_k^i - \phi(\delta_{k,k} + \delta_{k,k-1})$.

In the multigrid case, one can try to separate the computations at all levels $0, \dots, \ell$. We give a description of this algorithm, which looks rather similar to the multiplicative algorithm (27) with $\gamma = 1$. The following additive multigrid iteration $AMGM^{(0)}$ for solving $L_k u_k = f_k$ uses a damping factor ϕ , which is irrelevant if the iteration is embedded into a cg-like acceleration method.

```
function AMGM(k, u, f);
begin if k = 0 then AMGM :=  $\phi * L_0^{-1} * f$  else (44a)
  begin d := r (f - L_k * u); (44b)
    u := u +  $\phi * (S_k^*(u, f) - u)$ ; (44c)
    v := 0; v := AMGM(k-1, v, d); (44d)
    AMGM := u + p v (44e)
  end end;
```

Since (44b) does not influence the smoothing part (44c), both parts can be performed in parallel. The results of both parts are joined in (44e).

For the standard (multiplicative) two-grid method, we know from Section 2.3 that the convergence improves in

a particular way when the number v of smoothing steps increases. The same holds for general multigrid methods. However, this behavior is not true for the additive variant as shown by Bastian, Hackbusch and Witum (1998).

5.2 Interpretation as subspace iteration

Similar to algorithm (27), one can resolve the recursivity in (44e). For this purpose, we write explicitly $p = p_{k,k-1}$ and $r = r_{k,k-1}$ (as in (5), (6) and compose these mappings to obtain

$$p_{k,k} := p_{k,k-1} \circ p_{k-1,k-2} \circ \dots \circ p_{k+1,k}, \\ r_{k,k} := r_{k,k+1} \circ \dots \circ r_{k-1,k} \quad \text{for } k \leq \ell$$

where $p_{k,k} = r_{k,k} = I$ in the case of the empty product. From the given right-hand side f_k one defines $f_k := r_{k,k} f_k$ for $k = 0, \dots, \ell$. Then the additive multigrid iteration $AMGM^{(v)}(u_k, f_k)$ described in (44) is given by

$$AMGM^{(v)}(u_k, f_k) = u_k + \sum_{k=1}^{\ell} p_{k,k} \delta v_k \quad (45)$$

where the corrections δv_k at level k are defined by

$$\delta v_k := S_k^v(u_k, f_k) - u_k \quad \text{for } k = \ell \\ \delta v_k := S_k^v(0, f_k) \quad \text{for } k = 1, \dots, \ell - 1 \\ \delta v_0 := L_0^{-1} * f_0 \quad \text{for } k = 0$$

The case $k = \ell$, can easily be seen from (44c). Since the lower levels use the starting value $v = 0$, $S_k^v(u_k, f_k) - u_k$ simplifies to $S_k^v(0, f_k)$.

The interest in subspace iterations has two different reasons. (I) The computation of the corrections δv_k can be performed in parallel. (II) The interpretation as subspace iteration allows quite different convergence proofs (in particular for the V-cycle) than the standard multiplicative version. Although the resulting statements are weaker, they require also weaker assumptions (see Bramble and Zhang, 1993).

6 NESTED ITERATION

6.1 Algorithm

6.1.1 Starting and terminating an iteration

The natural approach is to start with a more or less accurate initial value u_k^0 and to perform several iteration steps:

$$\bar{u}_k := u_k^0; \text{ for } j := 1 \text{ to } i \text{ do } \bar{u}_k := MGM(\ell, \bar{u}_k, f_k) \quad (46)$$

The error of \bar{u}_k satisfies

$$\|\bar{u}_k - u_k\| \leq \zeta^i \|u_k^0 - u_k\| \quad (47)$$

where ζ is the contraction number of the iteration (cf. Section 2.3) and u_k the solution of $L_k u_k = f_k$. In particular, the simplest choice $u_k^0 = 0$ yields an estimate of the relative error

$$\frac{\|\bar{u}_k - u_k\|}{\|u_k\|} \leq \zeta^i \quad (\text{if } u_k^0 = 0) \quad (48)$$

In order to obtain a fixed (relative) error ε , one needs $i \geq \log(\varepsilon) / \log(\zeta) = O(|\log(\varepsilon)|)$ iterations, where we exploit the fact that the contraction number ζ of the multigrid iteration is ℓ -independent. Usually, ε is not explicitly given and one has to judge a suitable value. Except in special cases, it is useless to take ε smaller than the discretization error ε_{disc} (i.e. the difference between u_k and the continuous solution u). Often, the quantitative size of ε_{disc} is not known a priori, but only its asymptotic behavior $O(h_k^\alpha)$ (α : consistency order). From $\varepsilon := \varepsilon_{disc} = O(h_k^\alpha)$ one concludes that $i = O(|\log(h_k)|)$ iterations are required to obtain an iterate u_k with an error of the size of the discretization error. The corresponding number of arithmetical operations is $O(n_\ell |\log(h_k)|) = O(h_k^{-d} |\log(h_k)|)$.

6.1.2 Basic algorithm

The nested iteration described below (also called 'full multigrid method') has several advantages:

- Although no a priori knowledge of the discretization error is needed, the nested iteration produces approximations \bar{u}_k with error $O(\varepsilon_{disc})$.
- The nested iteration is cheaper than the simple approach (46). An approximation with error $O(\varepsilon_{disc})$ is calculated by $O(n_\ell)$ operations.
- Besides \bar{u}_k , the solutions $\bar{u}_{k-1}, \bar{u}_{k-2}, \dots$ corresponding to the coarser grids are also approximated and are at one's disposal.

In principle, the nested iteration can be combined with any iterative process. The idea is to provide a good starting guess u_k^0 by means of iterating on a coarser grid.

In this section, the index ℓ characterizes the level of the finest grid, that is, ℓ is the maximal level number, while the index $k \in \{0, 1, \dots, \ell\}$ is used for the intermediate levels. We assume that the discrete equations $L_k u_k = f_k$ are given for all levels.

A program-like formulation of the nested iteration reads as follows:

```

Nested Iteration
 $\bar{u}_0 := u_0 = L_0^{-1} f_0;$  (49a)
for  $k := 1$  to  $\ell$  do
  begin  $\bar{u}_k := \bar{p} \bar{u}_{k-1};$  (49b)
        for  $j := 1$  to  $i$  do  $\bar{u}_k := MGM(k, \bar{u}_k, f_k)$  (49c)
  end;

```

6.1.3 Implementational details

Starting Value at Level 0

The exact solution of $L_0 u_0 = f_0$ is not necessary. One may replace (49a) by $\bar{u}_0 \approx L_0^{-1} f_0$, provided that $\|\bar{u}_0 - u_0\|$ is small enough.

Prolongation \bar{p}

The starting value in (49b) is obtained from the coarse grid approximation \bar{u}_{k-1} by means of some interpolation \bar{p} . From the programming point of view, the simplest choice is $\bar{p} = p$ from (5). However, interpolations \bar{p} of higher order than p may be taken into considerations, too (see Remark 6).

If an asymptotic expansion is valid, Richardson's extrapolation can be applied to compute a fairly accurate value \bar{u}_k from \bar{u}_{k-1} and \bar{u}_{k-2} (details in Section 5.4 of Hackbusch, 1985).

Iterations per Level

At each level, i iterations are performed. An appropriate choice of i is discussed in Section 6.2. The same value i can be chosen for all levels, since the contraction numbers of the multigrid iteration are bounded independently of the level $0 \leq k \leq \ell$. Since most of the computational work is spent at level ℓ , it can be advantageous to choose $i_\ell \leq i_{\ell-1} = i_{\ell-2} = \dots = i_1$ (cf. Remark 8).

Adaptive Choice of the Finest Level

The nested iteration can be interpreted in two different ways.

From level ℓ down to 0. The finest level ℓ together with the stiffness matrix L_ℓ and the right-hand side f_ℓ is given. Then all levels $k < \ell$ are only introduced to support the multigrid process. The right-hand side f_k should be computed by $f_k := r_{k+1}^T f_{k+1}$ for $k = \ell - 1, \dots, 0$.

From level 0 to ℓ . The nested iteration becomes a part of the discretization process, if we choose the finer levels adaptively. Then the newest informations (a posteriori error estimates, comparison of \bar{u}_k and \bar{u}_{k-1} , etc.) can be used to decide whether a further refinement is needed.

In the second case, the loop $k = 1, \dots, \ell$ in (49) should be written as while-statement: ' $k := k + 1$, while the discretization is not fine enough'.

6.2 Analysis of the nested iteration

The nested iteration (49) requires the specification of an iteration number i . The following analysis suggests how to choose i .

Let ζ_k be the contraction number of the multigrid iteration employed at level k :

$$\|u_k^{j+1} - u_k\| \leq \zeta_k \|u_k^j - u_k\| \quad (50)$$

As pointed out before, the numbers ζ_k are uniformly bounded by some $\zeta < 1$. Set

$$\zeta := \max \{\zeta_k; 1 \leq k \leq \ell\} \quad (51)$$

where ℓ is the maximum level from (49). The discretization error is the (suitably defined) difference between the discrete solution $u_k = L_k^{-1} f_k$ and the continuous solution u . The difference between u_k and $u_{k-1} = L_{k-1}^{-1} f_{k-1}$ may be called the relative discretization error (error of u_{k-1} relative to u_k). We suppose an a priori estimate with a consistency order α ,

$$\|\bar{p} u_{k-1} - u_k\| \leq C_1 h_k^\alpha \quad \text{for } 1 \leq k \leq \ell \quad (52)$$

Note that the exponent α in (52) depends on the consistency order of the discretization and on the interpolation order of \bar{p} . Therefore, we are led to

Remark 6. The interpolation order of \bar{p} should at least be equal to the consistency order of the discretization.

Consider the standard case of a second order discretization ($\alpha = 2$) of a second order differential equation ($2m = 2$). By Note 6, \bar{p} should be at least piecewise linear. Hence, \bar{p} may be the standard prolongation p .

To indicate the levels involved, we write $\bar{p} = \bar{p}_{k+k-1}$. We define the constants

$$C_{20} := \max \{\|\bar{p}_{k+k-1}\|; 1 \leq k \leq \ell\} \quad (53a)$$

$$C_{21} := \max \left\{ \left(\frac{h_{k-1}}{h_k} \right)^\alpha; 1 \leq k \leq \ell \right\} \quad (53b)$$

$$C_2 := C_{20} C_{21} \quad (53c)$$

One can show $C_{20} = 1$ for the most frequent choices of \bar{p} . Moreover, $C_{21} = 2^\alpha$ holds for the usual sequence $h_k = h_0/2^k$. Hence, the value of C_2 is explicitly known: $C_2 = 2^\alpha$.

Theorem 2 (Error analysis) Assume (52) and $C_2 \zeta^i < 1$ with C_2 from (53a-c), ζ from (51), and i from (49). Set $C_3(\zeta, i) := \zeta^i / (1 - C_2 \zeta^i)$. Then the nested iteration (49) with i steps of the multigrid iteration per level results in \tilde{u}_k ($0 \leq k \leq \ell$) satisfying the error estimate

$$\|\tilde{u}_k - u_k\| \leq C_3(\zeta, i) C_1 h_k^i \quad \text{for all } k = 0, \dots, \ell \text{ with } u_k = L_k^{-1} f_k \quad (54)$$

Theorem 2 ensures that the errors at all levels $k = 0, 1, \dots, \ell$ differ from the (bound of the) relative discretization error $C_1 h_k^i$ only by a factor $C_3(\zeta, i)$, which is explicitly known. The only condition on i is $C_2 \zeta^i < 1$. The nested iteration is as cheap as possible if $i = 1$ satisfies $C_2 \zeta^i < 1$. With $C_2 = 2^*$ from above, this condition becomes $2^* \zeta < 1$. Assuming in addition the standard case of $\kappa = 2$, we obtain

Remark 7. Assume $C_2 = 4$ and $\zeta < 1/4$. Then Theorem 2 holds for $i = 1$. Note that the rates observed in Section 3.2 are far below the critical bound $1/4$.

Remark 8 (computational work) Let W_k be work for one multigrid iteration at level k . Assuming $h_{k-1} \approx 2h_k$ and $\Omega \subset \mathbb{R}^d$, the dimensions n_k of the system at level k should satisfy $n_k \approx 2^d n_{k-1}$. Since the work for \tilde{p} is less dominant, the cost of the nested iteration (49) is about $i * W_\ell / (1 - 2^{-d})$. In the 2D case, the value becomes $(4/3)iW_\ell$. Obviously, the complete nested iteration is only insignificantly more expensive than the work spent on level ℓ alone.

6.3 Numerical example

We apply the nested iteration (49) to the five-point scheme discretization of

$$\begin{aligned} -\Delta u &= f := -\Delta(\exp(x + y^2)) \quad \text{in } \Omega = (0, 1) \times (0, 1) \\ u &= \varphi := \exp(x + y^2) \quad \text{on } \Gamma = \partial\Omega \end{aligned} \quad (55)$$

The multigrid iteration MGM in (49) uses $v = \gamma = 2$. The results of the nested iteration are shown, where \tilde{p} is the cubic interpolation in the interior and the quadratic interpolation near the boundary. Let u_k^{exact} be the exact solution $\exp(x + y^2)$ restricted to the grid, while u_k is the exact discrete solution. We have to distinguish between the iteration errors $\|\tilde{u}_k - u_k\|$, the total error $\|\tilde{u}_k - u_k^{\text{exact}}\|$ with i from (49c), and the discretization error $\|u_k - u_k^{\text{exact}}\|$. The table shows the maximum norm of the total errors $\|\tilde{u}_k - u_k^{\text{exact}}\|$ for $i = 1$ and $i = 2$.

| ℓ | h_ℓ | $i = 1$ | $i = 2$ | $i = \infty$ |
|--------|----------|---------------------------|---------------------------|---------------------------|
| 0 | 1/2 | 7.9944658 ₁₀₋₃ | 7.9944658 ₁₀₋₃ | 7.9944658 ₁₀₋₃ |
| 1 | 1/4 | 3.9908756 ₁₀₋₃ | 2.9215605 ₁₀₋₃ | 2.8969488 ₁₀₋₃ |
| 2 | 1/8 | 1.5788721 ₁₀₋₃ | 8.1023136 ₁₀₋₄ | 8.0307789 ₁₀₋₄ |
| 3 | 1/16 | 3.2919346 ₁₀₋₄ | 2.0768391 ₁₀₋₄ | 2.0729855 ₁₀₋₄ |
| 4 | 1/32 | 5.7591549 ₁₀₋₄ | 5.2253758 ₁₀₋₄ | 5.2247399 ₁₀₋₄ |
| 5 | 1/64 | 1.3291689 ₁₀₋₄ | 1.3093946 ₁₀₋₄ | 1.3093956 ₁₀₋₄ |

The last column ($\tilde{u}_k^* = u_k$) contains the discretization error that should be in balance with the iteration error. Obviously, the choice $i = 1$ is sufficient. $i = 2$ needs the double work and cannot improve the total error substantially.

6.4 Use of acceleration by conjugate gradient methods

Given an iteration, it is often recommended to improve the convergence speed by the method of conjugate gradients (in the positive definite case) or by variants that apply to more general cases (see Hackbusch, 1994, Section 9). Here, two remarks are of interest.

If the matrix L_k in $L_k u_k = f_k$ is symmetric and positive definite, one should use a symmetric multigrid variant, that is, $MGM^{(v_1, v_2)}$ from Remark 2 with $v_1 = v_2$ and a symmetric smoothing iteration (it is even sufficient that pre-smoothing is adjoint to postsmoothing). Then the standard conjugate gradient (cg) method can be applied. However, the use of cg is recommended only if the rate of the multigrid convergence is not sufficiently fast; otherwise, the overhead for the cg-method does not pay.

7 NONLINEAR EQUATIONS

In the following, we consider the nonlinear elliptic problem $\mathcal{L}(u) = 0$, where \mathcal{L} is a nonlinear operator (e.g. $\mathcal{L}(u) = \operatorname{div} p(u) \operatorname{grad} u - f$ or $\mathcal{L}(u) = \Delta u + u u_x - f$). In the following, we assume that $\mathcal{L}(u) = 0$ is discretized with respect to a hierarchy of grids:

$$L_k(u_k) = f_k \quad \text{for } k = 0, \dots, \ell \quad (56)$$

Even if one is only interested in the solution of $L_k(u_k) = 0$, the multigrid approach in Section 7.2 leads to problems (56) with small right-hand sides f_k . The smallness of f_k ensures that (56) has a unique local solution (we do not require global uniqueness of the solutions).

7.1 Newton's method and linear multigrid

Newton's method requires the derivative (Jacobi matrix) of L_k , which we denote by $L_k(u_k) = (\partial/\partial u_k) L_k(u_k)$. Then the Newton iteration $u_k^{n+1} = u_k^n - [L_k(u_k^n)]^{-1} (L_k(u_k^n) - f_k)$ requires the solution of the linear system $L_k(u_k) \delta_k = d_k$ for the defect $d_k := L_k(u_k) - f_k$. The latter task can be performed by the multigrid iteration from above.

7.2 Nonlinear multigrid iteration

The approach from Section 7.1 requires the computation of the Jacobi matrix $L_k(u_k)$. This can be avoided by applying the nonlinear multigrid iteration $NMGM$ from (58) below. Since $NMGM$ uses approximations \tilde{u}_k of $L_k(u_k) = 0$ for $k \leq \ell - 1$, we start with the nonlinear nested iteration, which produces \tilde{u}_k as well as their defects $\tilde{f}_k := L_k(\tilde{u}_k)$.

```

solve  $L_0(\tilde{u}_0) = f_0$  approximately;
                                     (e.g. by Newton's method)
for  $k := 1$  to  $\ell$  do
  begin  $\tilde{f}_{k-1} := L_{k-1}(\tilde{u}_{k-1})$ ;          (defect of  $\tilde{u}_{k-1}$ )
        $\tilde{u}_k := \tilde{p} \tilde{u}_{k-1}$ ;                (start at level  $k$  as in (49b))
       for  $i := 1$  to  $i$  do  $\tilde{u}_k := NMGM(k, \tilde{u}_k, \tilde{f}_k)$ 
                                     (as in (49c))
  end;
                                     (57)

```

Now we define the iteration $NMGM$, which uses \tilde{u}_{k-1} , \tilde{f}_{k-1} as reference point (since in the linear case $\tilde{u}_{k-1} = 0$ is the reference point, we do not see \tilde{u}_{k-1} in the linear multigrid method).

```

function  $NMGM(k, u, f)$ ;
begin if  $k = 0$  then  $NMGM := \tilde{u}_0$  else
  ( $\tilde{u}_0$  approximation to  $L_0(\tilde{u}_0) = f$ )
  begin for  $i := 1$  to  $v$  do  $u := S_k(u, f)$ ;
        ( $\tilde{p}$ -smoothing)
         $d := r(L_k(u) - f)$ ;          (restriction of defect)
         $\varepsilon := \varepsilon(d)$ ;                  (small positive factor)
         $\delta := \tilde{f}_{k-1} - \varepsilon * d$ ; (right-hand side at level  $k - 1$ )
         $v := \tilde{u}_{k-1}$ ;                  (starting value for correction)
        for  $i := 1$  to  $\gamma$  do  $v := NMGM(k - 1, v, \delta)$ ;
         $NMGM := u + p(v - \tilde{u}_{k-1})/\varepsilon$ 
                                     (coarse grid correction)
  end end;
                                     (58)

```

Here, $S_k(u_k, f_k)$ is a nonlinear smoothing iteration for $L_k(u_k) = f_k$. For instance, the analogue of the Richardson iteration (4) is $S_k(u_k, f_k) = u_k - \omega_k(L_k(u_k) - f_k)$, where $\omega_k \approx 1/\|L_k(u_k)\|$. The factor $\varepsilon(d)$ may, for example, be chosen as $\sigma/\|d\|$ with a small number σ . The smallness of ε guarantees that $L_{k-1}(\tilde{u}_{k-1}) = \tilde{f}_{k-1} - \varepsilon * d$ has a unique

local solution close to \tilde{u}_{k-1} . Note that \tilde{f}_{k-1} , \tilde{u}_{k-1} are produced by (57).

Note that (58) is a true generalization of the linear iteration MGM : If $NMGM$ is applied to a linear problem (i.e. $L_k(u_k) := L_k u_k - f_k$), it produces the same iterates as MGM independently of the choice of the reference values \tilde{u}_{k-1} .

If $L_k(u_k)$ is Lipschitz continuous and if further technical conditions are fulfilled, one can show that the asymptotic convergence rate of $NMGM$ coincides with the rate of the linear iteration MGM applied to the linearized problem with the matrices $L_k := L_k(u_k^{\text{exact}})$. Also, other specifications of \tilde{u}_{k-1} , \tilde{f}_{k-1} , ε are possible (see Section 9 of Hackbusch, 1985).

If, by some reason, the nested iteration is not used, the value \tilde{u}_{k-1} can be replaced by u_k , while $\tilde{f}_{k-1} := L_{k-1}(\tilde{u}_{k-1})$ (FAS: 'full approximation storage method' from Brandt (1977)).

8 EIGENVALUE PROBLEMS

The (continuous) eigenvalue problem reads $Lu = \lambda u$, where u satisfies homogeneous boundary values. The hierarchy of discrete eigenvalue problems is

$$L_k u_k = \lambda u_k \quad \text{for } k = 0, \dots, \ell$$

(possibly with λI replaced by the mass matrix λM_k). Again, the two-grid method consists of smoothing and coarse grid correction based on the defect $d_k = L_k u_k - \lambda u_k$. However, the computation of a correction δu_k from $(L_k - \lambda I) \delta u_k = d_k$ is problematic, since $L_k - \lambda I$ becomes singular for the eigenvalue λ . Nevertheless, $(L_k - \lambda I) \delta u_k = d_k$ is solvable, since the right-hand side d_k belongs to the image space. Furthermore, the nonuniqueness of δu_k is harmless, since the kernel lies in the eigenspace. These statements are only approximately true for the restricted coarse grid equation $(L_{k-1} - \lambda I) v_{k-1} = r d_{k-1}$. Therefore, certain projections are necessary. The complete multigrid algorithms can be found in Section 12 of Hackbusch (1985).

If L_k is not symmetric, the right- and left-eigenvectors can be computed simultaneously. It is advantageous to compute a group of eigenvectors by combining the multigrid approach with the Ritz method (see Chapter 19, this Volume).

9 APPLICATIONS TO THE BOUNDARY ELEMENT METHOD (BEM)

There are two quite different groups of boundary element method (BEM) problems that can be solved by means of

multigrid methods. Integral equations of the second kind are treated in Section 9.1, while integral equations with hypersingular kernel are mentioned in Section 9.2. In both cases, the arising matrices K_k are fully populated. Except the coarsest grid, the only operation needed is the matrix-vector multiplication $u_k \mapsto K_k u_k$. Its naive implementation requires $O(n^2)$ arithmetical operations. Therefore, one should use the fast multiplication described in the author's contribution in Chapter 21, this Volume in this encyclopedia (see Chapter 12, this Volume).

9.1 Application to integral equations of the second kind

Fredholm integral equations of second kind have the form $\lambda u = Ku + f$ ($\lambda \neq 0$, f given) with

$$(Ku)(x) := \int_D s(x, y)u(y)dy \quad \text{for } x \in D$$

where the kernel s of the integral operator K is given. The Picard iteration $u \mapsto (1/\lambda)(Ku - f)$ converges only if $|\lambda| > \rho(K)$, but in many important applications the Picard iteration has a smoothing effect: nonsmooth functions e are mapped into a smooth function $(1/\lambda)Ke$ by only one step. This enables the following *multigrid iteration of the second kind*, which makes use of a hierarchy $\lambda_k u_k = K_k u_k + f_k$ of discrete equations.

function $MGM(k, u, f)$;
 MGM solving $\lambda_k u_k = K_k u_k + f_k$
begin if $k = 0$ **then** $MGM := (\lambda - K_0)^{-1}f$ **else**
 begin $u := \frac{1}{\lambda}(K_k * u + f)$; (Picard iteration)
 $d := r(\lambda_k u - K_k u - f_k)$;
 (restriction of defect)
 $v := 0$; (start value for correction)
 for $i := 1$ **to** 2 **do** $v := MGM(k-1, v, d)$;
 (W-cycle)
 $MGM := u - pv$ (coarse grid correction)
end end; (59)

Because of the strong smoothing effect, MGM has convergence rates $O(h_k^\alpha)$ with $\alpha > 0$. Hence, with increasing dimension (decreasing step size h_k) the iteration becomes faster! The value α depends on the smoothness of the image of K , on the order of the interpolation p , and so on.

Note that we need only the multiplication by the matrices K_k . It is not required that K is an integral operator with explicitly known kernel s . Iteration (59) applied to a fixed point equation $\lambda u = Ku + f$ has the same properties, provided that K shows the smoothing effect. For further details

we refer to Section 12 in Hackbusch (1985) and Section 5 in Hackbusch (1995).

The nonlinear fixed point equation $\lambda u = K(u)$ can be solved by an analogous version of the nonlinear multigrid method from Section 7.

9.2 Application to integral equations with hypersingular kernel

Boundary value problems $Lu = 0$ with inhomogeneous Neumann conditions can be formulated by means of hypersingular integral operators. For $L = \Delta$ and $\partial u / \partial n = \phi$ on $\Gamma = \partial \Omega \subset \mathbb{R}^3$, the solution u is given by the double-layer potential $u(x) = \int_\Gamma f(y)(\partial / \partial n_y)s(x, y) d\Gamma_y$ ($x \in \Gamma$) with $s(x, y) = 1/[4\pi\|x - y\|]$, provided that f satisfies

$$\int_\Gamma f(y) \frac{\partial}{\partial n_x} \frac{\partial}{\partial n_y} s(x, y) d\Gamma_y = \phi(x) \quad \text{for } x \in \Gamma.$$

Since $\partial^2 s / \partial n_x \partial n_y$ has a nonintegrable singularity, the integral must be understood in the sense of Hadamard. The variational formulation uses the energy space $\mathcal{H} = H^{1/2}(\Gamma)$. Using piecewise linear elements on Γ , we arrive at the setting (31) with the symmetric bilinear form

$$a(u, v) = \frac{1}{2} \int_\Gamma \int_\Gamma [u(x) - u(y)] \times [v(x) - v(y)] \frac{\partial}{\partial n_x} \frac{\partial}{\partial n_y} s(x, y) d\Gamma_x d\Gamma_y$$

(cf. Section 8.3 of Hackbusch, 1995). Since $a(\cdot, \cdot)$ is elliptic with respect to the energy space specified above, the discrete problems $A_k f_k = \phi_k$ can be solved in the same way as FEM systems from Section 4. In particular, p and r should be the canonical transfer mappings from Section 4.3. The only difference is that A_k is fully populated, whereas FEM matrices are sparse. This is the reason, why the fast panels clustering techniques should be implemented.

9.3 Application to first kind integral equations with weakly singular kernel

In the case of the single layer potential equation $Ku = f$ with $Ku = \int_\Gamma s(x, y)u(y) d\Gamma_y$ (an example of an integral equation of the first kind with weakly singular kernel) the standard smoothing procedure is not applicable. The reason is that the integral operator K has the negative order -1 (while the hypersingular integral operator from Section 9.2 has the positive order $+1$). As a consequence, the low eigenvalues of K are associated with the oscillatory

eigenfunctions, while the high eigenvalues belong to the smooth eigenfunctions. Therefore, a Richardson-like iteration reducing the high frequencies is not a smoothing procedure.

As a remedy (cf. Bramble, Leyk and Pasciak, 1993), the smoothing iteration must use a preconditioning of the equation by an operator D of order > 1 : $DKu = Df$ or $KDv = f$ or $D_1 K D_2 v = D_1 f$.

REFERENCES

- Bakhvalov NS. On the convergence of a relaxation method with natural constraints on the elliptic operator. *USSR Comput. Math. Math. Phys.* 1966; 6(5):101–135.
- Bastian P, Hackbusch W and Wittum G. Additive and multiplicative multi-grid – a comparison. *Computing* 1998; 60:345–364.
- Braess D. Towards algebraic multigrid for elliptic problems of second order. *Computing* 1995; 55:379–393.
- Braess D, Dryja M and Hackbusch W. Grid transfer for nonconforming FE-discretisations with application to non-matching grids. *Computing* 1999; 63:1–25.
- Brakhage H. Über die numerische Behandlung von Integralgleichungen nach der Quadratur-formel-methode. *Numer. Math.* 1960; 2:183–196.
- Bramble JH, Leyk Z and Pasciak JE. The analysis of multigrid methods for pseudo-differential operators of order minus one. *Math. Comp.* 1994; 63:461–478.
- Bramble JH and Zhang X. The analysis of multigrid methods. In *Volume VII of Handbook of Numerical Analysis*, Ciarlet PG, Lions JL (eds), Elsevier: Amsterdam, 1993; 173–415.
- Brandt A. Multi-level adaptive solutions to boundary-value problems. *Math. Comp.* 1977; 31:333–390.
- Brezina M, Mandel J and Vaneč P. Energy optimization of algebraic multigrid bases. *Computing* 1999; 62:205–228.
- Fedorenko RP. A relaxation method for solving elliptic difference equations. *USSR Comput. Math. Math. Phys.* 1961; 1:1092–1096.
- Fedorenko RP. The speed of convergence of one iterative process. *USSR Comput. Math. Math. Phys.* 1964; 4:227–235.

Haase G, Langer U, Reitzinger R and Schöberl J. Algebraic multigrid methods based on element preconditioning. *Int. J. Comput. Math.* 2001; 78:575–598.

Hackbusch W. *Multi-Grid Methods and Applications*, SCM 4. Springer: Berlin, 1985 (2nd printing) Springer: Berlin, 2003.

Hackbusch W. *Iterative Solution of Large Sparse Systems*. Springer: New York, 1994 – 2nd German edition: *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. Teubner: Stuttgart, 1993.

Hackbusch W. *Integral Equations. Theory and Numerical Treatment*, ISNM 128. Birkhäuser: Basel, 1995 – 2nd German edition: *Integralgleichungen. Theorie und Numerik*. Teubner: Stuttgart, 1997.

Hackbusch W and Sauter SA. Composite finite elements for the approximation of PDEs on domains with complicated microstructures. *Numer. Math.* 1997; 75:447–472.

Hackbusch W and Trottenberg U. *Multigrid Methods*, LNM 960. Springer: Berlin, 1982.

Hackbusch W and Trottenberg U. *Multigrid Methods II*, LNM 1228. Springer: Berlin, 1986.

Hackbusch W and Trottenberg U. *Multigrid Methods III*, ISNM 98. Birkhäuser: Basel, 1991.

Hackbusch W and Wittum G. *Multigrid Methods V*, LNCS 3. Springer: Berlin, 1998.

Hemker FW and Wesseling P. *Multigrid Methods IV*, ISNM 116. Birkhäuser: Basel, 1994.

Stüben K. Algebraic multigrid (AMG): experiences and comparisons. *Appl. Math. Comput.* 1983; 13:419–451.

Trottenberg U, Oosterlee C and Schüller A. *Multigrid*. Academic Press: San Diego, 2001.

Wesseling P. *An Introduction to Multigrid Methods*. Wiley: Chichester, 1991.

FURTHER READING

Hackbusch W. *Elliptic Differential Equations. Theory and Numerical Treatment*, SCM 18. Springer-Verlag: Berlin, 1992 (2nd printing, 2003) – 2nd German edition: *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner: Stuttgart, 1996.

Chapter 21

Panel Clustering Techniques and Hierarchical Matrices for BEM and FEM

Wolfgang Hackbusch

Max-Planck-Institut für Mathematik in den Naturwissenschaften, Inselstr., Leipzig, Germany

| | |
|--|-----|
| 1 Introduction | 597 |
| 2 The Panel Clustering Method (First Version) | 600 |
| 3 The Panel Clustering Method (Second Version) | 606 |
| 4 Hierarchical Matrices | 607 |
| References | 615 |

1 INTRODUCTION

The main background of the so-called 'panel clustering technique' is the efficient numerical treatment of *integral equations*. Therefore, we first refer the reader to the boundary element method (BEM) and the respective integral equations (see Section 1.2). The discrete problem is described by a fully populated $n \times n$ matrix. The naive approach requires a storage of the size n^2 and the standard matrix-vector multiplication needs $O(n^2)$ arithmetical operations. In order to realize the advantages of BEM compared with the finite element method (FEM), it is essential to reduce the order $O(n^2)$ of the cost to almost $O(n)$.

The panel clustering technique described in Section 2 allows reduction in the storage and matrix-vector costs

from $O(n^2)$ to $O(n \log^d n)$. The reduction in the memory is, in particular, important for 3D applications when $O(n \log^d n)$ data can easily be stored, while $O(n^2)$ exceeds the memory bounds. The reduction in the cost for the matrix-vector multiplication is important as well, since this is the essential operation in usual iterative methods for solving the system of linear equations. The essential ingredients of the panel clustering technique are (i) the far-field expansion (Section 2.1) and (ii) the panel cluster tree (Section 2.2). The chapter is concluded with some hints concerning implementational details (Section 2.7).

Section 3 presents a second variant of the panel clustering technique. This variant of the panel clustering technique can be generalized to the technique of hierarchical matrices (*H*-matrices), which is described in Section 4. Again, the *H*-matrix structure can be used to represent fully populated matrices. This technique allows not only the matrix-vector multiplication but also matrix operations like matrix-plus-matrix, matrix-times-matrix, and even matrix-inversion.

1.1 Notations

We have already used the Landau symbol $O(f(n))$, which means that the quantity is bounded by $C \cdot f(n)$ as $n \rightarrow \infty$ for some positive constant C . For an index set I , the set \mathbb{R}^I denotes the set of (real) vectors $\mathbf{a} = (a_i)_{i \in I}$ indexed by means of I . Similarly, the notation $\mathbb{R}^{I \times J}$ is used for the set of matrices $\mathbf{A} = (a_{i,j})_{i \in I, j \in J}$.

| Item | Explanation | References |
|--|--|---|
| A, B, \dots | Matrices of size $n \times n$ | (11) |
| b | Block, vertex of T_2 | Section 3.1.1 |
| b_j | BEM basis function | (10) |
| d | Spatial dimension of \mathbb{R}^d | (1) |
| diam, dist | Diameter and distance of clusters | (18), (38) |
| I | Index set for the matrix entries | Section 4.1 |
| I_m | Index set in the representation of $\tilde{\kappa}$ | (15) |
| $J_i^*, J_i^*(b_j)$ | Far-field coefficients | Section 2.4.3 |
| K | Integral operator | (4) |
| n | Problem dimension, matrix size | (10), Section 4.1 |
| \mathcal{P} | Set of panels (boundary elements) | Section 1.2.3 |
| $s(\mathbf{x}, \mathbf{y})$ | Fundamental solution | Section 1.2.1 |
| t | Triangle (panel) $t \in \mathcal{P}$ | Section 1.2.3 |
| $S, S_2, S_f, S_{f \times I}(\tau)$ | Set of sons | Section 2.2, Section 3.1.1, Section 4.2 |
| $T, T_2, T_1, T_{f \times I}$ | Cluster tree, tree of blocks, block cluster tree | Section 2.2, Section 3.1.1, Section 4.2 |
| \mathbf{u} | Coefficient vector from \mathbb{R}^d | (10) |
| V_h | Boundary element space | Section 1.2.3 |
| $\mathbf{x}, \mathbf{y}, \mathbf{z}$ | Points in \mathbb{R}^d | (2) |
| \mathbf{x}_τ | Center of cluster τ | Section 2.7.2 |
| Γ | Surface contained in \mathbb{R}^d | Section 1.2.2 |
| η | Parameter in admissibility condition | (18), (38) |
| $\kappa(\mathbf{x}, \mathbf{y})$ | Kernel of integral operator | (4) |
| $\tilde{\kappa}(\mathbf{x}, \mathbf{y}), \kappa_b(\mathbf{x}, \mathbf{y})$ | Far-field approximation of κ | (15), (40) |
| ξ, ξ^i | Collocation point | Section 1.2.3 |
| τ (also τ', σ, σ') | Cluster, vertex of the tree T | Section 2.2 |
| Φ, Φ_i | Expansion functions | (15) |
| $J_\Gamma, \dots, d\Gamma_x$ | Surface integration | (4) |
| $\#S$ | Cardinality of the set S , that is, number of elements | |

1.2 The boundary element method (BEM)

1.2.1 The problem to be solved

There are several important applications where an elliptic boundary value problem with a vanishing source term is to be solved,

$$Lu = 0 \quad \text{in } \Omega \subset \mathbb{R}^d \quad (1)$$

Here Ω may be a bounded or an unbounded domain. Since L is assumed to have constant coefficients, the fundamental solution $s(\mathbf{x}, \mathbf{y})$ is known explicitly. It satisfies $L_\tau s(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{x} - \mathbf{y})$, where $L_\tau = L$ is applied to the \mathbf{x} -argument and δ

is the Dirac function. In the case of $Lu = f \neq 0$, a further integral over Ω appears, which can be treated efficiently by means of the hierarchical matrices from Section 4. Examples of L and s are the Laplace problem,

$$L = \Delta, \quad s(\mathbf{x}, \mathbf{y}) = \begin{cases} \frac{1}{2\pi} \log |\mathbf{x} - \mathbf{y}| & \text{for } d = 2 \\ & (\text{i.e. } \mathbf{x}, \mathbf{y} \in \mathbb{R}^2) \\ \frac{1}{4\pi |\mathbf{x} - \mathbf{y}|} & \text{for } d = 3 \\ & (\text{i.e. } \mathbf{x}, \mathbf{y} \in \mathbb{R}^3) \end{cases} \quad (2)$$

the Helmholtz problem $L = \Delta + \sigma^2$, $s(\mathbf{x}, \mathbf{y}) = \exp(i\sigma |\mathbf{x} - \mathbf{y}|) / [4\pi |\mathbf{x} - \mathbf{y}|]$, and the Lamé equation ($d = 3$)

$$\mu \Delta \mathbf{u} + (\lambda + \mu) \nabla \operatorname{div} \mathbf{u} = 0$$

$$S(\mathbf{x}, \mathbf{y}) = \frac{\lambda + 3\mu}{8\pi(\lambda + 2\mu)} \times \left\{ \frac{1}{|\mathbf{x} - \mathbf{y}|} \mathbf{I} + \frac{\lambda + \mu}{\lambda + 3\mu} \frac{(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^\top}{|\mathbf{x} - \mathbf{y}|^3} \right\} \quad (3)$$

In the latter example, the fundamental solution $S(\mathbf{x}, \mathbf{y})$ is matrix-valued. In all examples, $|\mathbf{x} - \mathbf{y}|$ is the standard Euclidean norm of the vector $\mathbf{x} - \mathbf{y} \in \mathbb{R}^d$.

1.2.2 Formulation by an integral equation

The advantage of the following integral equation formulation is the fact that the domain of integration is the boundary $\Gamma = \partial\Omega$. Thus, the spatial dimension is reduced by one. This advantage is even more essential if Ω is an unbounded exterior domain.

There are several integral formulations based on integral operators K of the form

$$(Kf)(\mathbf{x}) := \int_\Gamma \kappa(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\Gamma_y \quad (4)$$

where $\kappa(\mathbf{x}, \mathbf{y})$ is $s(\mathbf{x}, \mathbf{y})$ or some derivative with respect to \mathbf{x} or \mathbf{y} . We give two examples.

Single-layer potential for a Dirichlet problem

Let the integral operator be defined by $\kappa = s$ with s from (2), that is, $(Kf)(\mathbf{x}) = (1/4\pi) \int_\Gamma f(\mathbf{y}) / |\mathbf{x} - \mathbf{y}| d\Gamma_y$ in the 3D case. Then $\Phi(\mathbf{x}) := (Kf)(\mathbf{x})$ is defined for all $\mathbf{x} \in \mathbb{R}^d$ and satisfies $\Delta\Phi = 0$ in $\mathbb{R}^d \setminus \Gamma$. In order to enforce the Dirichlet value,

$$\Phi = g \quad \text{on } \Gamma \quad (5)$$

the function f has to satisfy the integral equation

$$Kf = g \quad \text{for all } \mathbf{x} \in \Gamma$$

$$\text{that is, } \int_\Gamma \frac{f(\mathbf{y})}{|\mathbf{x} - \mathbf{y}|} d\Gamma_y = 4\pi g(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \Gamma \quad (6)$$

Therefore, one has to solve (a discrete version of) $Kf = g$. For the resulting solution f , the potential $\Phi = Kf$ fulfills (1) as well as (5) and can be evaluated at any point of interest.

Direct method

In (6), one has to solve for the unknown function f , which (indirectly) yields the solution of the Laplace problem after evaluation of $\Phi = Kf$. A direct approach is

$$\frac{1}{2} u(\mathbf{x}) = g(\mathbf{x}) + \int_\Gamma \kappa(\mathbf{x}, \mathbf{y}) u(\mathbf{y}) d\Gamma_y \quad \text{with } \kappa := \frac{\partial s}{\partial n_y}, \quad g(\mathbf{x}) := \int_\Gamma s(\mathbf{x}, \mathbf{y}) \phi(\mathbf{y}) d\Gamma_y \quad (7)$$

which yields the Dirichlet boundary values $u(\mathbf{x})$, $\mathbf{x} \in \Gamma$, of the interior domain with Neumann data ϕ . $s(\mathbf{x}, \mathbf{y})$ is the fundamental solution from (2). $\kappa(\mathbf{x}, \mathbf{y})$ is called the double-layer kernel. The equation on the left in (7) holds for almost all $\mathbf{x} \in \Gamma$ but must be corrected by a factor corresponding to the spherical angle of an edge or corner of the surface (cf. Hackbusch, 1997). This is important for the discretization by collocation but does not matter in the case of the Galerkin discretization.

1.2.3 Discretization by BEM

In the following, we assume the more interesting case of $d = 3$, that is, Γ is a two-dimensional surface.

Triangulation of the surface

To begin with, assume that the surface can be represented by a union of planar triangles: $\Gamma = \bigcup_{t \in \mathcal{P}} t$, where the triangulation \mathcal{P} is the set of these (closed) triangles. Usually, the triangulation is required to be conforming in the sense that the intersection of two different triangles is allowed to be either empty, a node, or an edge. Each $t \in \mathcal{P}$ can be produced by an affine map η_t from the unit triangle t_{unit} (vertices at $(0, 0)$, $(0, 1)$, $(1, 0)$) onto t , that is, $\eta_t(t_{\text{unit}}) = t$. In the following, we shall assume this simple case (of course, quadrilaterals instead of triangles are possible as well).

Alternatively, the true surface can be approximated by curved triangles, that is, Γ is replaced by $\bigcup_{t \in \mathcal{P}} \eta_t(t_{\text{unit}})$, where η_t is a more involved map producing a curved triangle.

In the BEM context, the triangles are often called panels.

Since the panels are assumed to be closed, two different panels may overlap at their boundaries. We say that the two subsets $s', s'' \subset \Gamma$ are weakly disjoint, if $\text{area}(s' \cap s'') = 0$. This covers the case of (completely) disjoint sets as well as the case when the boundaries overlap (but not the interior parts).

Boundary element space

The simplest boundary element is the piecewise constant one, that is, the boundary element space V_h that consists of functions being piecewise constant on each triangle $t \in \mathcal{P}$.

In the case of (continuous and) piecewise linear elements, the functions from V_h are (continuous and) piecewise affine on each triangle $t \in \mathcal{P}$. In the case of curved triangles, the piecewise affine functions on t_{unit} are mapped by η_t onto $\eta_t(t_{\text{unit}})$. Furthermore, one can consider spaces V_h of continuous or discontinuous functions that coincide with higher-order polynomials on $t \in \mathcal{P}$.

Galerkin discretization

The Galerkin discretization of $\lambda u + Ku = \phi$ with respect to the boundary element space V_h and K from (4) reads

$$\begin{aligned} \text{Find } u_h \in V_h \text{ such that} \\ \lambda \int_{\Gamma} u_h(x) v(x) d\Gamma_x + \int_{\Gamma} \int_{\Gamma} \kappa(x, y) u_h(y) v(x) d\Gamma_x d\Gamma_y \\ = \int_{\Gamma} \phi(x) v(x) d\Gamma_x \quad \text{for all } v \in V_h \end{aligned} \quad (8)$$

Collocation discretization

Since (8) involves a double integration, often the collocation is preferred, although the numerical statements about collocation are weaker. For this purpose, one defines a set $\Xi = \{\xi^i : i = 1, \dots, n\}$ of collocation points ξ^i , where $n = \dim V_h$. For instance, in the case of piecewise constant elements, $\xi \in \Xi$ should be chosen as the centroid of each $t \in \mathcal{P}$. Then, the collocation discretization of $\lambda u + Ku = \phi$ reads

$$\begin{aligned} \text{Find } u_h \in V_h \text{ such that} \\ \lambda u_h(\xi) + \int_{\Gamma} \kappa(\xi, y) u_h(y) d\Gamma_y = \phi(\xi) \quad \text{for all } \xi \in \Xi \end{aligned} \quad (9)$$

Matrix formulation

Let $\mathcal{B} = \{b_1, \dots, b_n\}$ be a basis of V_h . For instance, for piecewise constant elements, b_i is 1 on the i th triangle and 0 on each other $t \in \mathcal{P}$. In this case, we may use $t \in \mathcal{P}$ as an index instead of $i = 1, \dots, n$, that is, the basis is $\mathcal{B} = \{b_t : t \in \mathcal{P}\}$.

In the case of discontinuous and piecewise linear elements, we have three basis functions per triangle: $\mathcal{B} = \{b_{t,k} : t \in \mathcal{P}, k = 1, 2, 3\}$, while for continuous and piecewise linear elements, each basis function b_t is associated with a vertex x_t of the triangulation.

Each $u_h \in V_h$ is represented by

$$u_h = \sum_{j=1}^n u_j b_j \quad (10)$$

where $\mathbf{u} = (u_j)_{j=1, \dots, n}$ abbreviates the coefficient vector.

Then the solution of the collocation problem (9) is characterized by

$$\lambda \mathbf{A} \mathbf{u} + \mathbf{B} \mathbf{u} = \mathbf{f} \quad (11)$$

where the matrices \mathbf{A} and \mathbf{B} and the vector \mathbf{f} are given by

$$\begin{aligned} \mathbf{A} &= (b_j(\xi^i))_{i,j=1, \dots, n} \\ \mathbf{B} &= \left(\int_{\Gamma} \kappa(\xi^i, y) b_j(y) d\Gamma_y \right)_{i,j=1, \dots, n} \\ \mathbf{f} &= (\phi(\xi^i))_{i=1, \dots, n} \end{aligned} \quad (12)$$

The Galerkin solution is given by (10) and (11) with

$$\begin{aligned} \mathbf{A} &= \left(\int_{\Gamma} b_j(x) b_i(x) d\Gamma_x \right)_{i,j=1, \dots, n} \\ \mathbf{B} &= \left(\int_{\Gamma} \int_{\Gamma} \kappa(x, y) b_j(y) b_i(x) d\Gamma_x d\Gamma_y \right)_{i,j=1, \dots, n} \\ \mathbf{f} &= \left(\int_{\Gamma} \phi(x) b_i(x) d\Gamma_x \right)_{i=1, \dots, n} \end{aligned} \quad (13)$$

In the case of (12), $\mathbf{A} = \mathbf{I}$ holds, provided that b_j is the Lagrange function. In any case, \mathbf{A} is a sparse matrix, which causes no problems. Differently, \mathbf{B} is usually a *fully populated* matrix. Standard representation needs a storage of n^2 for all entries. The panel clustering method will reduce this size to $\mathcal{O}(n \log^2 n)$, that is, the storage will be almost linear in the dimension n . The same improvement holds for the cost of the matrix-vector multiplication.

For further details about integral equations and boundary elements, see Hackbusch (1997) (see Chapter 12, this Volume).

2 THE PANEL CLUSTERING METHOD (FIRST VERSION)

The panel clustering method was introduced in the eighties (cf. Hackbusch and Nowak, 1986). The multipole method, which started at the same time (cf. Greengard and Rokhlin, 1997), is similar, with the difference that it is designed more for point charges and requires an operator-dependent construction for the expansion functions. Quite another, but theoretically related, approach is the matrix compression, which can be applied in the case of a proper wavelet discretization (cf. Dahmen, Prössdorf and Schneider, 1993).

The first version, which we present now, corresponds to the collocation equation (11) and more precisely to the performance of the matrix-vector multiplication by \mathbf{B} , that is, $\mathbf{u} \mapsto \mathbf{B} \mathbf{u}$. We recall that the i th component of $\mathbf{B} \mathbf{u}$ reads

$$(\mathbf{B} \mathbf{u})_i = \sum_{j=1}^n u_j \int_{\Gamma} \kappa(\xi^i, y) b_j(y) d\Gamma_y \quad (14)$$

(see (12)). For the fast evaluation of this term, we have to introduce the far-field expansion (Section 2.1) and the panel cluster tree (Section 2.2).

2.1 Far-field expansion

Consider one of the collocation points $\mathbf{x} = \xi^i$ and approximate the kernel $\kappa(\mathbf{x}, \mathbf{y})$ in a subset $\tau \subset \Gamma \setminus \{\mathbf{x}\}$ by a finite expansion of the form

$$\tilde{\kappa}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{l} \in I_m} c_{\mathbf{l}}^{\tau}(\mathbf{x}) \Phi_{\mathbf{l}}(\mathbf{y}) \quad (15)$$

where $\Phi_{\mathbf{l}}(\mathbf{y})$ for $\mathbf{l} \in I_m$ are functions that are independent of \mathbf{x} . I_m is an index set whose size is indicated by m (see (16) below). The upper index τ denotes a subset of Γ in which $\tilde{\kappa}(\mathbf{x}, \cdot)$ should be a good approximation of $\kappa(\mathbf{x}, \cdot)$.

In the simplest case, $\tilde{\kappa}(\mathbf{x}, \cdot)$ is the Taylor expansion around the center of τ up to degree $m-1$. In this case, the index set I_m equals the set $\{\mathbf{l} \in \mathbb{N}_0^d : l_1 + \dots + l_d \leq m-1\}$ of multi-indices, where d is the spatial dimension. In order to make the functions $\Phi_{\mathbf{l}}(\mathbf{y})$ independent of τ , we may choose the monomials $\Phi_{\mathbf{l}}(\mathbf{y}) = y_1^{l_1} \times \dots \times y_d^{l_d}$ (expansion around $\mathbf{z}(\tau)$). In the case discussed above, the number of indices of I_m is bounded by (16) with $C_l = 1$:

$$\#I_m \leq C_l m^d \quad (16)$$

Further details about the approximation of κ by (15) will follow in Section 2.7.2.

2.2 Cluster tree

Assembling several neighboring triangles (panels) $t \in \mathcal{P}$, we can form *clusters*. The union of neighboring clusters yield even larger clusters. This process can be continued until the complete surface Γ is obtained as the largest cluster. As in Figure 1, the clusters may have an irregular geometric shape (they may even be unconnected). For later purpose, it is favorable if clusters are rather compact, that is, $\text{area}(c)/\text{diam}(c)^2$ should be large.

One may consider this process also from the opposite point of view (domain decomposition). The surface Γ is divided into smaller parts (clusters) that are divided further until only the panels remain as the trivial clusters. For a construction based on this approach, see Section 2.7.1.

The process of clustering (or repeated domain decomposition) is represented by the *cluster tree* T . In the following definition, \mathcal{P} is the set of panels, and we denote the set of unions of panels by $\mathcal{S} = \{\bigcup_{t \in \tau} t : \tau \subset \mathcal{P}\}$. All panels $t \in \mathcal{P}$ belong to \mathcal{S} but also $\Gamma = \bigcup_{t \in \tau} t \in \mathcal{S}$.

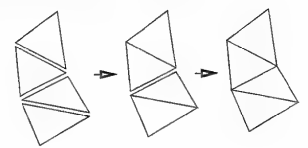


Figure 1. Clustering of four triangles.

Definition 1. (a) All vertices of T belong to \mathcal{S} . (b) $\Gamma \in T$ is the root of T . (c) The leaves of T are the panels from \mathcal{P} . (d) If $\tau \in T$ is no leaf, there is a set $S(\tau)$ with at least two sons, which are weakly disjoint (cf. Section 1.2.3). Furthermore, $S(\tau)$ satisfies

$$\tau = \bigcup_{\tau' \in S(\tau)} \tau' \quad (17)$$

Usually, $S(\tau)$ consists of exactly two sons so that T becomes a binary tree.

Remark 1. The number of clusters $\tau \in T$ is at most $\#T \leq 2\#\mathcal{P} - 1 = 2n - 1$. The upper bound $\#T = 2n - 1$ holds for a binary tree.

2.3 Admissible clusters and admissible coverings

For the integration of $\kappa(\mathbf{x}, \mathbf{y}) u(\mathbf{y})$ over a cluster τ (cf. (14)), we shall use the expansion $\tilde{\kappa}(\mathbf{x}, \mathbf{y})$ from (15) instead of κ . This requires the following condition on \mathbf{x} and τ .

We call $\tau \in T$ to be an *admissible cluster* with respect to some control point $\mathbf{x} \in \mathbb{R}^d$ if

$$\text{diam}(\tau) \leq \eta \text{dist}(\mathbf{x}, \tau) \quad (18)$$

The parameter $\eta > 0$ will be chosen later (η will turn out to be constant, independent of the panel size h). From inequality (18), we see that the larger the distance between \mathbf{x} and the cluster, the larger the cluster may be.

A set of clusters $\mathcal{C} = \{\tau_1, \dots, \tau_s\} \subset T$ is called a *covering* (of Γ) if the clusters are weakly disjoint and satisfy

$$\Gamma = \bigcup_{j=1}^s \tau_j \quad (19)$$

There are two trivial coverings; $\mathcal{C} = \{\Gamma\}$ is the coarsest one. The finest is $\mathcal{C} = \mathcal{P}$. In the first case, the cluster is as large as possible (but the number of clusters is minimum);

in the second case, the clusters are as small as possible (but their number is at a maximum).

In the following, we are looking for a covering that (from the computational point of view) should consist of a small number of clusters and that, on the other hand, should be admissible. This leads us to the following definition.

Definition 2. We call $C = \{\tau_1, \dots, \tau_n\} \subset T$ an *admissible covering* (of Γ) with respect to \mathbf{x} if it is a covering satisfying (19) and

$$\text{either } \tau_j \in \mathcal{P} \text{ or } \tau_j \text{ is admissible with respect to } \mathbf{x} \quad (20)$$

Remark 2. (a) If $\mathbf{x} \in \Gamma$, there is no covering (19) of Γ consisting of only admissible clusters. (b) Condition (20) states that inadmissible clusters are panels.

The number of clusters in C should be as small as possible. The optimum is discussed in

Proposition 1. For each $\mathbf{x} \in \mathbb{R}^d$, there is a unique admissible covering $C(\mathbf{x})$ with respect to \mathbf{x} with minimum number $n_C(\mathbf{x}) := \#C(\mathbf{x})$ of clusters. $C(\mathbf{x})$ is called the minimum admissible covering with respect to \mathbf{x} .

The minimum admissible covering $C(\mathbf{x})$ with respect to \mathbf{x} can be easily computed by

$$C := \emptyset; \text{ Divide}(\Gamma, C); \text{ comment the result is } C = C(\mathbf{x}) \quad (21a)$$

where *Divide* is the recursive procedure

```

procedure Divide( $\tau, C$ ); comment  $\tau \in T$  is a cluster,
                         $C$  is a subset of  $T$ ;
begin if  $\tau$  is admissible with respect to  $\mathbf{x}$  then
                         $C := C \cup \{\tau\}$ 
    else if  $\tau \in \mathcal{P}$  then  $C := C \cup \{\tau\}$ 
    else for all  $\tau' \in S(\tau)$  do Divide( $\tau', C$ )
end;

```

(21b)

2.4 Algorithm for the matrix-vector multiplication

2.4.1 Partition into near and far field

Instead of computing the matrix entries, we compute the far-field coefficients in Phase I. In Phase II, we evaluate an approximation of Ku_k at $\mathbf{x} \in \mathbb{R}^d$ (e.g. at $\mathbf{x} = \xi^i \in \mathbb{Z}$). Repeating (14), we recall that the desired result of the

matrix-vector multiplication $\mathbf{v} = \mathbf{B}\mathbf{u}$ is

$$v_i = \sum_{j=1}^n u_j \int_{\Gamma} \kappa(\xi^i, \mathbf{y}) b_j(\mathbf{y}) d\Gamma_{\mathbf{y}} \quad (22)$$

Let $C(\xi^i)$ be the minimum admissible covering determined in (21b). We split $C(\xi^i)$ into a *near-field* and a *far-field* part defined by

$$C_{\text{near}}(\xi^i) := \{\tau \in C(\xi^i) : \tau \text{ is not admissible}\}$$

$$C_{\text{far}}(\xi^i) := \{\tau \in C(\xi^i) : \tau \text{ is admissible}\}$$

All $\tau \in C_{\text{near}}(\xi^i)$ are panels (cf. Remark 2b). The integral in (22) can be written as

$$\int_{\Gamma} \dots = \sum_{\tau \in C_{\text{near}}(\xi^i)} \int_{\tau} \dots + \sum_{\tau \in C_{\text{far}}(\xi^i)} \int_{\tau} \dots$$

This induces an analogous splitting of v_i from (22) into

$$v_i = v_i^{\text{near}} + v_i^{\text{far}}$$

The part v_i^{far} will be approximated by \tilde{v}_i^{far} in Section 2.4.3. Note that the splitting depends on i . Another i' yields another collocation point $\xi^{i'}$ and another splitting into $C_{\text{near}}(\xi^{i'})$ and $C_{\text{far}}(\xi^{i'})$.

2.4.2 Near-field part

The near-field part of v_i is computed exactly (or with sufficiently accurate quadrature):

$$v_i^{\text{near}} = \sum_{\tau \in C_{\text{near}}(\xi^i)} \sum_j u_j \int_{\tau} \kappa(\xi^i, \mathbf{y}) b_j(\mathbf{y}) d\Gamma_{\mathbf{y}} \quad (23)$$

Since the support of the basis functions b_j is small, there are only a constant number of indices j such that a panel $\tau \in C_{\text{near}}(\xi^i)$ intersects with $\text{supp}(b_j)$. Hence, the sum \sum_j has only $\mathcal{O}(1)$ terms. The number of panels in $C_{\text{near}}(\xi^i)$ turns out to be bounded by $\mathcal{O}(\log n)$.

2.4.3 Far-field part

Replacing the exact kernel κ in the definition of v_i^{far} by $\tilde{\kappa}(\mathbf{x}, \mathbf{y})$ from (15), we obtain

$$\tilde{v}_i^{\text{far}} := \sum_{\tau \in C_{\text{far}}(\xi^i)} \sum_j u_j \int_{\tau} \tilde{\kappa}(\xi^i, \mathbf{y}) b_j(\mathbf{y}) d\Gamma_{\mathbf{y}}$$

$$= \sum_{\tau \in C_{\text{far}}(\xi^i)} \sum_j u_j \int_{\tau} \sum_{i' \in I_m} \kappa_i^{i'}(\xi^i) \Phi_{i'}(\mathbf{y}) b_j(\mathbf{y}) d\Gamma_{\mathbf{y}}$$

Summation and integration can be interchanged

$$\tilde{v}_i^{\text{far}} = \sum_{i' \in I_m} \sum_{\tau \in C_{\text{far}}(\xi^i)} \kappa_i^{i'}(\xi^i) \sum_j u_j \int_{\tau} \Phi_{i'}(\mathbf{y}) b_j(\mathbf{y}) d\Gamma_{\mathbf{y}} \quad (24)$$

The following integrals are called *far-field coefficients*:

$$J_i^i(b_j) := \int_{\tau} \Phi_{i'}(\mathbf{y}) b_j(\mathbf{y}) d\Gamma_{\mathbf{y}}, \quad (i \in I_m, \tau \in T, 1 \leq j \leq n) \quad (25)$$

Of particular interest are the far-field coefficients corresponding to panels. These are the only ones to be evaluated in the first phase:

$$J_i^i(b_j) := \int_{\tau} \Phi_{i'}(\mathbf{y}) b_j(\mathbf{y}) d\Gamma_{\mathbf{y}}, \quad (i \in I_m, \tau \in \mathcal{P}, 1 \leq j \leq n) \quad (26)$$

Remark 3. (a) The coefficients $J_i^i(b_j)$ are independent of the special vector \mathbf{u} in the matrix-vector multiplication $\mathbf{v} = \mathbf{B}\mathbf{u}$. (b) There are only a fixed number of panels τ intersecting with the support of b_j ; otherwise, $J_i^i(b_j) = 0$. The number of nonzero coefficients $J_i^i(b_j)$ is $\mathcal{O}(n \#I_m)$.

As soon as the far-field coefficients (26) are computed, the quantities

$$J_i^i := \sum_j u_j J_i^i(b_j) \quad (i \in I_m, i \in \mathcal{P}) \quad (27)$$

can be summed up by $\mathcal{O}(n \#I_m)$ additions. For $\tau \in T \setminus \mathcal{P}$, we exploit the tree structure:

$$J_i^i = \sum_{\tau \in S(\tau)} J_i^i \quad \text{for } \tau \in T \setminus \mathcal{P} \quad (28)$$

The coefficients J_i^i represent the sum

$$J_i^i = \sum_j u_j J_i^i(b_j) = \int_{\tau} \Phi_{i'}(\mathbf{y}) \sum_j u_j b_j(\mathbf{y}) d\Gamma_{\mathbf{y}}$$

Hence, the quantities \tilde{v}_i^{far} can be computed from the simple sum

$$\tilde{v}_i^{\text{far}} = \sum_{\tau \in C_{\text{far}}(\xi^i)} \sum_{i' \in I_m} \kappa_i^{i'}(\xi^i) J_i^i \quad (29)$$

(see (24)). Since the number of clusters $\tau \in C_{\text{far}}(\xi^i)$ is expected to be much smaller than the number n of all panels, representation (29) should be advantageous.

2.5 The additional quadrature error

The replacement of v_i^{far} by \tilde{v}_i^{far} can be regarded as an additional quadrature error. The error of the expansion (15)

depends on the order m and the cluster containing \mathbf{y} . The exact requirements on $\tilde{\kappa}$ are as follows.

Assumption 1. Let $\eta_0 \in (0, 1)$ and a ball $B \subset \mathbb{R}^d$ be given. There are constants C_1 and C_2 such that for all $0 < \eta < \eta_0 < 1$ and $m \in \mathbb{N}$, there are expansions $\tilde{\kappa}$ of the form (15) satisfying

$$|\kappa(\mathbf{x}, \mathbf{y}) - \tilde{\kappa}(\mathbf{x}, \mathbf{y})| \leq C_1 (C_2 \eta)^m |\kappa(\mathbf{x}, \mathbf{y})|$$

$$\text{for all } \mathbf{y} \in B \text{ and } \text{diam } B \leq \eta \text{dist}(\mathbf{x}, B) \quad (30)$$

Inequality (30) provides an estimation of the relative error of $\tilde{\kappa}$. The proof of (30) in Hackbusch and Nowak (1989) uses a Taylor expansion $\tilde{\kappa}$ with respect to \mathbf{y} for standard examples and determines the values of C_1 , C_2 , and η_0 . The estimation in (30) by $\kappa(\mathbf{x}, \mathbf{y})$ on the right-hand side makes, for instance, sense for positive kernels like $\kappa(\mathbf{x}, \mathbf{y}) = (1/4\pi |\mathbf{x} - \mathbf{y}|)$. In Section 2.7.3, it will become obvious that (30) can also be obtained for the double-layer kernel (see (7)), although its sign changes.

By (30), the error is of order $\mathcal{O}(\eta^m)$. In order to make the error equal to the consistency error $\mathcal{O}(h^s)$ (h : panel size), we choose η in (18) by

$$\eta = \eta(n) := \mathcal{O}((n^{-s/(d-1)})^{1/m}) < \eta_0, \quad \kappa < m \quad (31)$$

Then $\mathcal{O}(h^m)$ equals $\mathcal{O}(h^s)$. Since, in (33), m will be chosen as $\mathcal{O}(\log n)$, the quantity η becomes independent of h .

2.6 Complexity of the algorithm

2.6.1 Choice of parameters

The complexity of the algorithm depends mainly on the number $n_C(\mathbf{x})$ of clusters in the minimum admissible covering $C(\mathbf{x})$. Under natural conditions described in Hackbusch and Nowak (1989), where also details of the proofs can be found, there is a constant C_C with

$$n_C(\mathbf{x}) \leq n_C(\eta, n) := C_C \left(\frac{1}{\eta} \right)^{d-1} \log(2 + \eta^{d-1} \#P)$$

$$\text{for all } \mathbf{x} \in \mathbb{R}^d \quad (32)$$

with η from (18). The logarithmic factor can even be omitted if $\mathbf{x} \notin \Gamma$.

Inserting (31) into (32) and using $\#P = \mathcal{O}(n)$, we obtain the following estimate for the number $n_C(\mathbf{x})$ of clusters in $C(\mathbf{x})$:

$$n_C(\mathbf{x}) \leq n_C(\eta, n) = C_C n^{s/m} \log(2 + C_d n^{1-s/m})$$

$$\text{for all } \mathbf{x} \in \mathbb{R}^d$$

The optimal choice of the expansion order m turns out to be

$$m := \left\lfloor \frac{x}{d+1} \log n \right\rfloor \quad (|x| := \text{largest integer } i \text{ with } i \leq x) \quad (33)$$

Then $n^{x/m}$ (and η) is a constant, and we obtain the estimate

$$n_C(x) \leq C \log n \quad (34)$$

Therefore, we have to deal with only $O(\log n)$ clusters instead of n panels.

2.6.2 Operation count

While Phase I has to be performed only once for initialization, Phase II has to be repeated for every matrix-vector multiplication.

Phase I (a) Algorithm (21a, b) (computing the minimum admissible covering $C(x)$) requires $O(n_C(x))$ operations per point x . Because of (34), and since there are n different collocation points $x = \xi^i$, the total amount of work in this part is $O(n \log n)$.

(b) The computation of v_i^{nm} in (23) needs $O(\log n)$ evaluations of integrals of the form $\int_{\Gamma} \kappa(\xi^i, y) b_j(y) d\Gamma_y$ per index i , that is, in total $O(n \log n)$ evaluations.

(c) The far-field coefficients $J_i^*(b_j)$ ($i \in \mathcal{P}$, $i \in I_m$) can be computed by $O(n \log^d n)$ operations (cf. (16), (33)) and require $O(n \log^d n)$ evaluations (or approximations) of integrals of the form $\int_{\Gamma} \Phi_i(y) b_j(y) d\Gamma_y$.

(d) The number of coefficients $\kappa_i^*(\xi^i)$ to be evaluated for $\tau \in C_{\text{int}}(\xi^i)$, $i \in I_m$, $1 \leq i \leq n$ equals $O(n \log^{d-1} n)$.

Phase II (a) The far-field coefficients J_i^* for the nontrivial clusters $\tau \in T \setminus \mathcal{P}$ and all indices $i \in I_m$ can be summed up in $O(n \log^d n)$ additions (see (28)).

(b) The final summation in (29) requires only $O(n \log^{d+1} n)$ additions.

Theorem 1. (a) The data computed in Phase I and the quantities from Phase II require a storage of size $O(n \log^{d+1} n)$ data. (b) Each matrix-vector multiplication $u \mapsto Bu$ can be approximated up to an error of size $O(h^*)$ by $O(n \log^{d+1} n)$ operations.

Concerning the storage in Phase I, we remark that only the coverings $C(\xi^i)$, the nonzero integrals $\int_{\Gamma} \kappa(\xi^i, y) b_j(y) d\Gamma_y$ from (23), the expansion coefficients $\kappa_i^*(\xi^i)$, and the far-field coefficients $J_i^*(b_j)$ are to be stored. The costs for Phase I can be further reduced if several panels are geometrically similar.

2.7 Some implementational details

2.7.1 Construction of the cluster tree

In the following, we describe the construction of the cluster tree T by means of bounding boxes. This method is, in particular, suited for elements in a domain $\Omega \subset \mathbb{R}^d$. The application to surfaces will be discussed at the end of this section.

Associate every element t with its centroid denoted by z_t , with the coordinates $z_{t,k}$ ($k = 1, \dots, d$). Let Z be the set of these points. The smallest bounding box $Q \subset \mathbb{R}^d$ containing all z_t is given by

$$Q = [a_1, b_1] \times \dots \times [a_d, b_d], \quad \text{where } a_k := \min\{z_{t,k} : z_t \in Z\}, \quad b_k := \max\{z_{t,k} : z_t \in Z\} \quad (35)$$

Choose $k \in \{1, \dots, d\}$ such that the side length $b_k - a_k$ is maximum and divide Q into

$$Q^I = [a_1, b_1] \times \dots \times [a_k, a_k + \frac{b_k - a_k}{2}] \times \dots \times [a_d, b_d] \\ Q^{II} = [a_1, b_1] \times \dots \times [a_k + \frac{b_k - a_k}{2}, b_k] \times \dots \times [a_d, b_d]$$

This gives rise to a partition of Z into $Z^I := Z \cap Q^I$ and $Z^{II} := Z \cap Q^{II}$. The procedure can be repeated recursively: Determine the bounding box Q' of Z' (it may be smaller than Q^I !) and split Q' into Q'^I and Q'^{II} and accordingly Z' . The recursion stops when the resulting subset of Z contains only one point. Obviously, the construction produces a binary tree T_Z starting with the root Z . Any vertex of T_Z is a subset Z' of Z . Since each $z \in Z'$ corresponds to exactly one panel $t \in \mathcal{P}$, the union $\bigcup_{z \in Z'} t_z$ describes a cluster. In this way, the tree T_Z can be transferred into the desired cluster tree.

Figure 2 shows the bisection process in the two-dimensional case.

For BEM, there is a modification that is of interest. As soon as the corresponding cluster is close to some (tangent) hyperplane, the coordinates of the bounding box can be rotated so that $d-1$ coordinates are in the hyperplane, while the d th coordinate is in the normal direction.

2.7.2 Far-field expansion by polynomial interpolation

In (15), $\kappa(x, y) = \sum_{i \in I_m} \kappa_i^*(x) \Phi_i(y)$ describes the approximation of $\kappa(x, y)$ in the cluster τ for a fixed collocation point x . Let $d_m = \#I_m$ denote the dimension of polynomials of total degree $m-1$. Choose d_m interpolation points ζ_i^* and let $\tilde{\kappa}(x, y)$ be the polynomial interpolating $\kappa(x, y)$. It has the

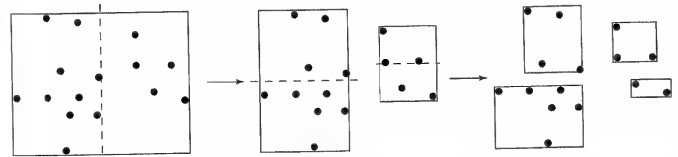


Figure 2. The bounding box to the left containing the points z_t is divided in two parts in z_1 -direction. In the second step, the new bounding boxes are divided in z_2 -direction.

representation $\sum_{i=1}^{d_m} \kappa(x, \zeta_i^*) L_i^*(y)$, where $L_i^*(y)$ denotes the Lagrange polynomials with the property $L_i^*(\zeta_j^*) = \delta_{ij}$ (Kronecker symbol). Expanding L_i^* into monomials, one obtains the desired representation (15).

Concerning the choice of interpolation points, one is not restricted to the cluster $\tau \subset \Gamma \subset \mathbb{R}^d$. Instead, one can first define suitable quadrature points ζ_i in the unit cube $C = [-1/2, 1/2]^d$. Given a cluster τ with center z_τ , consider the cube $C_\tau := z_\tau + \text{diam}(\tau)C$ and use the corresponding quadrature points $\zeta_i^* = z_\tau + \text{diam}(\tau)\zeta_i$. The interpolation polynomial converges to the Taylor polynomial if all interpolation points ζ_i^* tend to the center z_τ .

The latter approach requires that $\kappa(x, \cdot)$ is defined on C_τ . This holds, for example, for kernels from Section 1.2.1 but not for kernels involving normal derivatives as the double-layer kernel, since the normal derivative is defined on Γ only. The remedy is given in the next subsection.

2.7.3 Far-field expansion for kernels with normal derivatives

The double-layer kernel for the Laplace problem is $(\partial/\partial n_y)(1/4\pi|x-y|) = (1/4\pi)((x-y, n(y))/|x-y|^2)$ (cf. (7)). One possibility is to approximate $1/|x-y|^2$ by an expression of the form $\sum_{i=1}^d \kappa_i^*(x) \Phi_i(y)$. Then $(1/4\pi)((x-y, n(y))/|x-y|^2)$ is approximated by $\sum_{i=1}^d \{x_j \kappa_i^*(x)\} [n_j(y) \Phi_i(y)] - \sum_{i=1}^d \kappa_i^*(x) \{y, n(y)\} \Phi_i(y)$ and the latter expression is again of the form (15). Note that nonsmooth surfaces yielding nonsmooth normal directions $n(y)$ cause no difficulty. Furthermore, the relative error estimate (30) can be shown (the error becomes zero if $(1/4\pi)((x-y, n(y))/|x-y|^2) = 0$ due to $x-y \perp n(y)$).

The disadvantage of the described approach is the fact that the number of terms is multiplied by the factor 4. This can be avoided by approximating $(1/4\pi|x-y|)$ by $\sum_{i=1}^d \kappa_i(x) \Phi_i^*(y)$ and forming its normal derivative: $\sum_{i=1}^d \kappa_i(x) (\partial/\partial n_y) \Phi_i^*(y)$, which gives (15) with $\Phi_i(y) := (\partial/\partial n_y) \Phi_i^*(y)$.

The latter approach is, in particular, helpful when the Lamé equation is treated. Note that $((\partial/\partial x_i)(\partial/\partial x_j)|x-y|)_{i,j=1,\dots,3} = (1/|x-y|)I - ((x-y)(x-y)/|x-y|^3)$; hence, the expansion of the scalar function $|x-y|$ has to be differentiated.

2.8 Modification: approximations with basis transforms

In $\tilde{\kappa}(x, y) = \sum_{i \in I_m} \kappa_i^*(x) \Phi_i(y)$ (cf. (15)), we required $\Phi_i(y)$ to be independent of τ . This fact was used in (28): the quadrature results of $J_i^*(b_j) = \int_{\Gamma} \Phi_i(y) b_j(y) d\Gamma_y$ for the sons $\tau' \in S(\tau)$ could be used to get $J_i^*(b_j) = \int_{\Gamma} \Phi_i(y) b_j(y) d\Gamma_y$, as their sum.

However, a global basis $\{\Phi_i(y) : i \in I_m\}$ has numerical disadvantages. Considering, for example, polynomials, one likes to have locally defined bases $\{\Phi_i(y) : i \in I_m\}$ for each cluster $\tau \in T$. Since these different bases span the same spaces, there are transformations of the form

$$\Phi_i^*(b_j) = \sum_{\lambda \in I_m} \omega_{i,\lambda}^* \Phi_\lambda^*(y) \quad \text{for } \tau \in T \text{ and } \tau' \in S(\tau) \quad (36)$$

We redefine

$$J_i^*(b_j) := \int_{\Gamma} \Phi_i^*(y) b_j(y) d\Gamma_y \quad (37)$$

using the τ -dependent basis $\{\Phi_i^*(y) : i \in I_m\}$. The computation starts at the leaves (panels): $J_i^*(b_j)$ is computed for all $i \in \mathcal{P}$. Owing to (36), we have $J_i^*(b_j) = \sum_{\tau' \in S(\tau)} \sum_{\lambda \in I_m} \omega_{i,\lambda}^* J_\lambda^*(b_j)$ instead of (25). We store only $J_i^*(b_j)$ for $i \in \mathcal{P}$ and compute for a given vector u the quantities $J_i^*(b_j) = \sum_{j \in \mathcal{P}} u_j J_i^*(b_j)$ as in (27). However, the formula for $J_i^* = \sum_{j \in \mathcal{P}} u_j J_i^*(b_j)$ has now to use (37) and reads

$$J_i^* = \sum_{\tau' \in S(\tau)} \sum_{\lambda \in I_m} \omega_{i,\lambda}^* J_\lambda^*$$

instead of (28). These J_i^* can now be used in (29) to obtain \tilde{v}_i^{far} .

Concerning the coefficients $\omega_{\tau, \sigma}^{\tau, \sigma}$, we return to the basis of Lagrange functions L_i^* defined in Section 2.7.2. In that case, $\omega_{\tau, \sigma}^{\tau, \sigma} = L_i^*(\xi_i^*)$ involves nothing more than the evaluation of the τ -basis functions at the interpolation points associated to τ .

Another obvious basis are the monomials $(y - z_\tau)^i$ centered around the midpoint z_τ of the cluster τ . In this case, (36) describes the re-expansion of polynomials centered at z_τ around the new center z_σ .

3 THE PANEL CLUSTERING METHOD (SECOND VERSION)

The previous version of the panel clustering method is completely row-oriented. For each row index i , we compute the component v_i of $v = Bu$ by means of a covering $C(\xi_i^*)$ that may change with i . As a consequence, the kernel $\kappa(x, y) = \kappa(\xi_i^*, y)$ is a function of y only and (15) describes an expansion with respect to y .

In the following, we try to determine a version in which the x - and y -directions are treated equally. This is, in particular, more appropriate for the Galerkin discretization (8).

The tree T from Section 2.2 was introduced to describe decompositions of Γ . Now we consider the product $\Gamma \times \Gamma$ and determine a corresponding (second) tree T_2 , in Section 3.1.1. The vertices of T_2 are products $\tau \times \sigma \subset \Gamma \times \Gamma$ of clusters $\tau, \sigma \in T$. The kernel $\kappa(x, y)$ will be approximated by a special separable expansion (40) for $(x, y) \in \tau \times \sigma$.

3.1 The tree T_2 of products of clusters

3.1.1 Definition

Let the cluster tree T be defined as in Section 2.2. The second tree T_2 is constructed from T as follows. We use the symbol b for the vertices of T_2 . While $S(\tau)$ denotes the set of sons of $\tau \in T$, the sons of $b \in T_2$ form the set $S_2(b)$.

Definition 3. (a) T_2 is a subset of $T \times T$, that is, each vertex of T_2 is a product $\tau \times \sigma$ of two clusters $\tau, \sigma \in T$. (b) $\Gamma \times \Gamma \in T_2$ is the root of the tree. (c) It remains to construct the set $S_2(b)$ of sons of any $b = \tau \times \sigma \in T_2$:

$$S_2(b) := \begin{cases} \{\tau' \times \sigma' : \sigma' \in S(\sigma), \text{ if neither } \sigma \text{ nor } \tau \\ \quad \tau' \in S(\tau) \} & \text{are leaves of } T_2 \\ \{\tau' \times \sigma' : \sigma' \in S(\sigma) \text{ if } \tau \text{ is a leaf of } T_2 \text{ but not } \sigma \\ \quad \{\tau' \times \sigma' : \tau' \in S(\tau) \text{ if } \sigma \text{ is a leaf of } T_2 \text{ but not } \tau \\ \quad \emptyset & \text{if } \tau \text{ and } \sigma \text{ are leaves of } T_2 \end{cases}$$

The last case, $S_2(b) = \emptyset$ is equivalent to saying that b is a leaf in T_2 . Note that (b) defines the first vertex in T_2 , while by (c), one recursively gets new vertices belonging to T_2 . In this way, T_2 is completely defined by T . In particular, only the tree structure of T has to be stored.

Remark 4. The tree T_2 has the same properties as T in (17): For any $b \in T_2$ not being a leaf, the sons $b' \in S_2(b)$ are weakly disjoint and $b = \bigcup_{b' \in S_2(b)} b'$. The leaves of T_2 are of the form $\tau \times \sigma$ with panels $\tau, \sigma \in \mathcal{P}$.

3.1.2 Admissibility, covering C_2

Let $\eta > 0$ be the same parameter as in (18). A product $b = \tau \times \sigma \in T_2$ is called *admissible* if

$$\max(\text{diam}(\tau), \text{diam}(\sigma)) \leq \eta \text{dist}(\tau, \sigma) \quad (38)$$

As in Definition 2, we define a covering of $\Gamma \times \Gamma$.

Definition 4. (a) A covering $C_2 \subset T_2$ is a subset with pairwise weakly disjoint $b \in C_2$ such that $\bigcup_{b \in C_2} b = \Gamma \times \Gamma$. (b) An *admissible covering* C_2 is a covering such that all $b \in C_2$ are either admissible or are leaves.

Again, we are looking for a minimum admissible covering C_2 , which is obtained by (39a) using *Divide2* from (39b).

$$C_2 := \emptyset; \text{Divide2}(\Gamma \times \Gamma, C_2); \quad (39a)$$

procedure *Divide2*(b, C_2); **comment** $b \in T_2, C_2 \subset T_2$;

begin if b is admissible then $C_2 := C_2 \cup \{b\}$

else if b is a leaf of T_2 then $C_2 := C_2 \cup \{b\}$

else for all $b' \in S_2(b)$ do *Divide2*(b', C_2)

end; (39b)

In the following, C_2 denotes the minimum admissible covering obtained by (39a-b).

3.2 Kernel expansion

We split C_2 into a far field $C_2^{\text{far}} := \{b \in C_2 : b \text{ is admissible}\}$ and near field $C_2^{\text{near}} := \{b \in C_2 : b \text{ is not admissible}\}$. In the latter case, b is a leaf of T_2 . Owing to the admissibility condition (38), the kernel function $\kappa(x, y)$ allows an expansion with respect to x and $y \in \sigma$ when $b \in C_2^{\text{far}}$. For this purpose, we introduce a basis $\{\Phi_\tau^* : \tau \in T_\sigma\}$ for each cluster $\tau \in T$, which is applied with respect to x and y .

Given $b = \tau \times \sigma \in C_2^{\text{far}}$, we approximate $\kappa(x, y)$ by an expression $\kappa_b(x, y)$ of the form

$$\kappa_b(x, y) := \sum_{v \in I_\sigma} \sum_{\mu \in I_\sigma} \kappa_{v, \mu}^b \Phi_v^*(x) \Phi_\mu^*(y) \quad (40)$$

for $(x, y) \in b = \tau \times \sigma \in C_2^{\text{far}}$

An example of such an expression is the Taylor expansion with respect to (x, y) around the centers (z_τ, z_σ) of τ and σ . Then $\Phi_v^*(x)$ is the monomial $(x - z_\tau)^v$, where $v \in I_\sigma$ belongs to the same set of multi-indices as for the Taylor expansion in Section 2.1.

The coefficients $\kappa_{v, \mu}^b$ in (40) form a $d_\sigma \times d_\sigma$ matrix $K^b = (\kappa_{v, \mu}^b)_{v, \mu \in I_\sigma}$, where $d_\sigma := \#I_\sigma$.

3.3 Matrix-vector multiplication for the Galerkin discretization

We recall the Galerkin discretization (8) and the matrix formulation (13) involving the matrix B . The i th component of $v = Bu$ is

$$v_i = \sum_{j=1}^n u_j \int_\Gamma \int_\Gamma \kappa(x, y) b_j(y) b_i(x) d\Gamma_x d\Gamma_y$$

The covering C_2 allows to replace the integration over $\Gamma \times \Gamma$ by the sum of integrals over $b \in C_2$

$$v_i = \sum_{b \in C_2} u_j \int_b \int_b \kappa(x, y) b_j(y) b_i(x) d\Gamma_x d\Gamma_y$$

If $b = \tau \times \sigma \in C_2^{\text{near}}$, the expression remains unchanged and yields the near-field part v_i^{near} . For $b = \tau \times \sigma \in C_2^{\text{far}}$, we replace $\kappa(x, y)$ by $\kappa_b(x, y)$ from (40),

$$\begin{aligned} v_i^{\text{far}} &= \sum_{b \in C_2^{\text{far}}} \sum_{j=1}^n u_j \int_b \int_b \sum_{v \in I_\sigma} \sum_{\mu \in I_\sigma} \kappa_{v, \mu}^b \Phi_v^*(x) \Phi_\mu^*(y) \\ &\quad \times b_j(y) b_i(x) d\Gamma_x d\Gamma_y \\ &= \sum_{b \in C_2^{\text{far}}} \sum_{j=1}^n u_j \sum_{v \in I_\sigma} \sum_{\mu \in I_\sigma} \kappa_{v, \mu}^b \left(\int_\tau \Phi_\tau^*(x) b_i(x) d\Gamma_x \right) \\ &\quad \times \left(\int_\sigma \Phi_\sigma^*(y) b_j(y) d\Gamma_y \right) \\ &= \sum_{b \in C_2^{\text{far}}} \sum_{j=1}^n u_j \sum_{v \in I_\sigma} \sum_{\mu \in I_\sigma} \kappa_{v, \mu}^b J_\tau^*(b_i) J_\sigma^*(b_j) \end{aligned}$$

$$= \sum_{b \in C_2^{\text{far}}} \sum_{v \in I_\sigma} \sum_{\mu \in I_\sigma} \kappa_{v, \mu}^b J_\tau^*(b_i) J_\sigma^*(b_j)$$

with quantities $J_\sigma^*(b_j)$ and $J_\tau^*(b_i)$ already defined in Section 2.8.

4 HIERARCHICAL MATRICES

The panel clustering method was element oriented and enabled a fast matrix-vector multiplication. The present method is index oriented and supports all matrix operations, that is, additionally an approximate addition, multiplication, and inversion of matrices are possible.

The technique of hierarchical matrices (\mathcal{H} -matrices) applies not only to full BEM matrices, but also to the fully populated inverse stiffness matrices of FEM problems.

Again, the construction is based on trees that are similar to those from Section 3. However, the panels are replaced by the indices, that is, by the degrees of freedom. This will lead to a block-structured matrix, where all subblocks are filled with low-rank matrices.

The use of low-rank matrices for subblocks was already proposed by Tyrtyshnikov (1996); however, the construction of the efficient block-structure was missing.

4.1 Index set I

We consider square matrices $A = (a_{ij})_{i, j \in I}$, where the indices i, j run through the index set I of size $n := \#I$. We shall not use an explicit naming of the indices by $I = \{1, \dots, n\}$, since this might lead to the wrong impression that the indices must have a special ordering. The technique of \mathcal{H} -matrices can easily be extended to rectangular matrices $B = (b_{ij})_{i \in I, j \in J}$, where I and J are different index sets.

For the following construction, we need to know some geometric information about the indices. The simplest case is given by point data:

Assumption 2. Each index $i \in I$ is associated with a 'nodal point' $\xi^i \in \mathbb{R}^d$.

In this case, we use the following obvious definitions for the diameter of a subset $I' \subset I$ and for the distance of two subsets $I', I'' \subset I$:

$$\begin{aligned} \text{diam}(I') &:= \max \{ \|\xi^i - \xi^j\| : i, j \in I' \} \\ \text{dist}(I', I'') &:= \min \{ \|\xi^i - \xi^j\| : i \in I', j \in I'' \} \end{aligned} \quad (41a)$$

where $\|\cdot\|$ denotes the Euclidean norm in \mathbb{R}^d .

Although this information is sufficient for the practical application, precise statements (and proofs) about the FEM (or BEM) Galerkin method require the following support information:

Assumption 3. Each index $i \in I$ is associated with the support $X_i := \text{supp}(b_i) \subset \mathbb{R}^d$ of the finite element basis function b_i . For subsets $I', I'' \subset I$ we define

$$\begin{aligned} X(I') &:= \bigcup_{i \in I'} X_i \\ \text{diam}(I') &:= \max\{|x - y| : x, y \in X(I')\} \\ \text{dist}(I', I'') &:= \min\{|x - y| : x \in X(I'), y \in X(I'')\} \end{aligned} \quad (41b)$$

4.2 Cluster tree T_I for \mathcal{H} -matrices

The following tree T_I is constructed as the panel cluster tree T from Definition 1, but the panels $i \in P$ are replaced by indices $i \in I$.

Definition 5. The cluster tree T_I consisting of subsets of I is structured as follows.

- (a) $I \in T_I$ is the root of T_I .
- (b) The leaves of T_I are given by the one-element sets $\{i\}$ for all $i \in I$.
- (c) If $\tau \in T_I$ is no leaf, there exist disjoint sons $\tau_1, \dots, \tau_k \in T_I$ ($k = k(\tau) > 1$) with $\tau = \tau_1 \cup \dots \cup \tau_k$.

We denote the set of sons by $S_I(\tau)$. Usually, binary trees ($k = 2$) are appropriate.

Remark 5. (a) The fact that the leaves of T_I contain exactly one element is assumed in order to simplify the considerations. In practice, one fixes a number C_{leaf} (e.g. $C_{\text{leaf}} = 32$) and deletes all $\tau' \in S_I(\tau)$ with $\#\tau' \leq C_{\text{leaf}}$, that is, in the reduced tree, the leaves are characterized by $\#\tau \leq C_{\text{leaf}}$. Definition 5 corresponds to $C_{\text{leaf}} = 1$.

(b) The construction from Section 2.7.1 can be used as well to build T_I . The centers \mathbf{x}_i from Section 2.7.1 (cf. Figure 2) are to be replaced by the points ξ_i^l from Assumption 2.

4.3 Block cluster tree $T_{I \times I}$

The entries a_{ij} of a matrix $A \in \mathbb{R}^{I \times I}$ are indexed by pairs $(i, j) \in I \times I$. Accordingly, the block cluster tree $T_{I \times I}$ contains subsets of $I \times I$. Given the tree T_I , the block cluster tree $T_{I \times I}$ is constructed similar to T_I in Section 3.1.1. The vertices (blocks) of $T_{I \times I}$ are denoted

by b ; the sons of b form the set $S_{I \times I}(b)$. For a matrix $A \in \mathbb{R}^{I \times I}$ and a block $b \in T_{I \times I}$, the corresponding submatrix is denoted by

$$A|_b := (a_{i,j})_{(i,j) \in b} \quad (42)$$

Definition 6. (a) The vertices of $T_{I \times I}$ are products $b = \tau \times \sigma$ of two clusters $\tau, \sigma \in T_I$.

(b) $I \times I \in T_{I \times I}$ is the root of the tree.

(c) The set of sons of $b = \tau \times \sigma \in T_{I \times I}$ is defined by

$$S_{I \times I}(b) := \{\tau' \times \sigma' : \tau' \in S(\tau), \sigma' \in S(\sigma)\}$$

Note that $S_{I \times I}(b) = \emptyset$ if either $S(\tau)$ or $S(\sigma)$ is the empty set. Hence, the leaves of $S_{I \times I}(b)$ are those $b = \tau \times \sigma$ where either τ or σ are leaves of T_I .

4.4 Admissibility condition

Next, we need an *admissibility condition* that allows us to check if a block b is of appropriate size. We recall that $\text{diam}(\sigma)$ and $\text{dist}(\tau, \sigma)$ ($\tau, \sigma \in T_I$) are defined by (41a) or (b).

Definition 7. Let $\eta > 0$ be a fixed parameter. The block $b = \tau \times \sigma \in T_{I \times I}$ is called *admissible*, if either b is a leaf or

$$\min\{\text{diam}(\tau), \text{diam}(\sigma)\} \leq \eta \text{dist}(\tau, \sigma) \quad (43)$$

Note that in (43), the *minimum* of the diameters appears, while the panel clustering in (38) needs the maximum.

The simplest way to check admissibility condition (43) is to apply (43) to the bounding boxes Q_σ and Q_τ from (35). The condition $\min\{\text{diam}(Q_\sigma), \text{diam}(Q_\tau)\} \leq 2\eta \text{dist}(Q_\sigma, Q_\tau)$, which is easy to verify, implies (43).

4.5 Admissible block partitioning

The first step in the construction of \mathcal{H} -matrices is the block partitioning (see, e.g. Figure 3). The partitioning called P is a covering in the sense that all blocks $b \in P$ are disjoint and $\bigcup_{b \in P} b = I \times I$. The partitioning is admissible if all $b \in P$ are admissible in the sense of Definition 7. Again, we are looking for a minimum admissible partitioning for which $\#P$ is as small as possible. It can be determined as in (21a,b) or (39a,b). We apply (44) with DivideP from (44),

$$P := \emptyset; \text{DivideP}(I \times I, P); \quad (44a)$$

procedure $\text{DivideP}(b, P)$; comment $b \in T_{I \times I}$, $P \subset T_{I \times I}$;

begin if b is admissible then $P := P \cup \{b\}$

else if b is a leaf of $T_{I \times I}$ then $P := P \cup \{b\}$

else for all $b' \in S_{I \times I}(b)$ do $\text{DivideP}(b', P)$

end; (44b)

Next, we give an example of such a minimum admissible partitioning that corresponds to a discretization of the integral operator $\int_0^1 \log|x - y| f(y) dy$, where $d = 1$ is the spatial dimension. Consider the piecewise constant boundary elements that give rise to the supports

$$\begin{aligned} X_i &:= [(i-1)h, ih] \quad \text{for } i \in I := \{1, \dots, n\} \\ \text{and } h &:= \frac{1}{n}, \text{ where } n = 2^p \end{aligned}$$

(cf. (41)). The cluster tree T_I is the binary tree obtained by a uniform halving: the resulting clusters form the tree $T_I = \{\tau_i^l : 0 \leq \ell \leq p, 1 \leq i \leq 2^\ell\}$, where

$$\tau_i^l = \{(i-1) * 2^{p-\ell} + 1, (i-1) * 2^{p-\ell} + 2, \dots, i * 2^{p-\ell}\} \quad (45)$$

Note that $\tau_1^0 = I$ is the root, while $\tau_i^p = \{i\}$ are the leaves. Further, we choose $\eta = 1$ in (43). Then the resulting block partitioning is shown in Figure 3. Under natural conditions, the number of blocks is $\#P = \mathcal{O}(n)$.

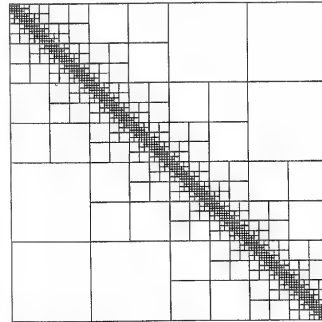


Figure 3. Partitioning for a 1D example.

4.6 \mathcal{H} -matrices and R_k -matrices

Definition 8 (\mathcal{H} -matrix) Given a cluster tree T_I for an index set I , let the related minimum admissible partitioning be denoted by P . Further, let $k \in \mathbb{N}$ be a given integer. Then the set $\mathcal{H}(k, P)$ consists of all matrices $M \in \mathbb{R}^{I \times I}$ with

$$\text{rank}(M|_b) \leq k \quad \text{for all } b \in P$$

Any rectangular matrix $C \in \mathbb{R}^b$ ($b = \tau \times \sigma$) with $\text{rank}(C) \leq k$ gives rise to the following equivalent representations

$$\begin{aligned} C &= \sum_{i=1}^k \mathbf{a}_i \mathbf{b}_i^T \in \mathbb{R}^b \quad \text{with vectors } \mathbf{a}_i \in \mathbb{R}^\tau, \mathbf{b}_i \in \mathbb{R}^\sigma, \\ C &= \mathbf{A} \mathbf{B}^T \quad \text{with } \mathbf{A} \in \mathbb{R}^{\tau \times k}, \mathbf{B} \in \mathbb{R}^{\sigma \times k} \end{aligned} \quad (46)$$

where the matrices $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_k]$, $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_k]$ are composed by the vectors $\mathbf{a}_i, \mathbf{b}_i$. The vectors in (46) may be linearly dependent, since $\text{rank}(M) < k$ is not excluded. Throughout this section, the bound k on the rank is assumed to be much smaller than the dimension n .

Definition 9 (R_k -matrix) Matrices represented in the form (46) are called R_k -matrices.

Remark 6. (a) R_k -matrices require a storage of size $2k(\#\tau + \#\sigma)$.

(b) Multiplication of an R_k -matrix $C = \mathbf{A} \mathbf{B}^T$ with a vector requires k scalar products and vector additions: $C \mathbf{v} = \sum_{i=1}^k \alpha_i \mathbf{a}_i$ with $\alpha_i := (\mathbf{b}_i, \mathbf{v})$. The cost is $2k(\#\tau + \#\sigma)$.

(c) Multiplication of two R_k -matrices $\mathbf{R} = \mathbf{A} \mathbf{B}^T \in \mathbb{R}^b$, $\mathbf{S} = \mathbf{C} \mathbf{D}^T \in \mathbb{R}^{b'}$ with $b = \tau \times \sigma$, $b' = \sigma' \times \sigma''$ leads to the R_k -matrix $\mathbf{T} = \mathbf{E} \mathbf{D}^T \in \mathbb{R}^{\tau \times \sigma''}$ with $\mathbf{E} = \mathbf{A} * \mathbf{Z}$, where $\mathbf{Z} = \mathbf{B}^T \mathbf{C}$ is of size $k \times k$. The operation cost is $2k^2(\#\tau + \#\sigma)$.

(d) The product $\mathbf{M} \mathbf{R}$ for an arbitrary matrix $\mathbf{M} \in \mathbb{R}^b$ and an R_k -matrix $\mathbf{R} \in \mathbb{R}^b$ is again an R_k -matrix of the form (46) with $\mathbf{a}_i' := \mathbf{M} \mathbf{a}_i$.

According to Remark 5, the leaves of T_I may be characterized by $\#\tau \leq C_{\text{leaf}}$. Then all submatrices $M|_b$ of an \mathcal{H} -matrix M are represented as R_k -matrices except for the case when $b = \tau \times \sigma$ with $\#\sigma, \#\tau \leq C_{\text{leaf}}$, where a (usual) full matrix is preferable.

Remark 7. Under rather general assumptions on the tree T_I and on the geometric data ξ_i^l or X_i (cf. Section 4.1), the storage requirements for any $M \in \mathcal{H}(k, P)$ are $\mathcal{O}(nk \log(n))$ (cf. Hackbusch, 1999; Hackbusch and Khoromskij, 2000).

The constant in the estimate $O(nk \log(n))$ from Remark 7 is determined by a *sparsity constant* C_m of the partitioning P (see Grasedyck and Hackbusch, 2003).

4.7 Hierarchical format for BEM and FEM matrices

So far, the set $\mathcal{H}(k, P)$ of matrices is defined. It is still to be shown that matrices of this format are able to approximate well with those matrices that we want to represent.

4.7.1 BEM matrices

The second version of the panel clustering method is already quite close to the present form. In the former case, the data associated with a block $b = \tau \times \sigma \in T_2$ (notation in the sense of Section 3.1.1) describe the part $\int_{\tau} \int_{\sigma} \kappa(x, y) b_1(x) u(y) dx dy$ with κ approximated by $\tilde{\kappa}$. In the special case $u = b_j$ (i.e. u is the j th unit vector), this integral becomes $\int_{\tau} \int_{\sigma} \kappa(x, y) b_1(x) b_j(y) dx dy$. Now, we want to represent $a_{ij} = \int_{\tau} \int_{\sigma} \kappa(x, y) b_i(x) b_j(y) dx dy$ which, in general, is different from the previous result, since the supports X_i, X_j of b_i, b_j are not necessarily contained in τ and σ . Owing to the admissibility of the block $b = \tau \times \sigma \in P$ (notation in the sense of Section 4.2), $\kappa(x, y)$ is approximated by

$$\tilde{\kappa}(x, y) = \sum_{l=1}^k \Psi_l(x) \Phi_l(y) \quad \text{for all } x \in X(\tau), y \in X(\sigma) \quad (47)$$

(cf. (15)). Inserting one term of (47) into $\int_{\tau} \int_{\sigma} \dots b_1(x) b_j(y) dx dy$ for $(i, j) \in b$, we obtain $\alpha_i * \beta_j$, where $\alpha_i := \int_{\tau} \Psi_l(x) b_i(x) dx$ and $\beta_j := \int_{\sigma} \Phi_l(y) b_j(y) dy$. These components form the vectors $\mathbf{a}_i = (\alpha_i)_{i \in \tau}$ and $\mathbf{b}_j = (\beta_j)_{j \in \sigma}$. Hence,

$$\int_{\tau} \int_{\sigma} \tilde{\kappa}(x, y) b_i(x) b_j(y) dx dy = \sum_{l=1}^k \mathbf{a}_l \mathbf{b}_l^T \quad \text{for } (i, j) \in b$$

shows that the approximation of κ in $X(\tau) \times X(\sigma)$ by $\tilde{\kappa}$ containing k terms is equivalent to having an Rk -submatrix $A|_b$ with $\text{rank}(A|_b) \leq k$. In the case of (admissible) blocks $b \in P$, which do not satisfy (43), $b = \{(i, j)\}$ is of size 1×1 (cf. Definition 7), so that a_{ij} can be defined by the exact entry.

Remark 8. (a) Let $A \in \mathbb{R}^{I \times I}$ be the exact BEM matrix. The existence of a (well approximating) \mathcal{H} -matrix $\tilde{A} \in \mathcal{H}(k, P)$ follows from a sufficiently accurate expansion (47) with k terms for all 'far-field blocks' $b \in P$ satisfying (43).

(b) The BEM kernels (mathematically more precisely: asymptotically smooth kernels; cf. Hackbusch and Khoromskij, 2000) allow an approximation up to an error of $O(\eta^{d/(d-1/4)})$ by k terms (η is the factor in (43)).

Concerning the construction of $\tilde{A} \in \mathcal{H}(k, P)$, one can follow the pattern of panel clustering (see, e.g. Börm and Hackbusch, 2002 and Börm, Grasedyck and Hackbusch, 2003). Interestingly, there is another approach called *adaptive cross approximation* (ACA) by Bebendorf (2000); Bebendorf and Rjasanov (2003), which only makes use of the procedure $(i, j) \mapsto \int_{\tau} \int_{\sigma} \kappa(x, y) b_i(x) b_j(y) dx dy$ (this mapping is evaluated only for a few index pairs $(i, j) \in I \times I$).

4.7.2 FEM matrices

Since FEM matrices are sparse, we have the following trivial statement.

Remark 9. Let (41) be used to define the admissible partitioning P . Then, for any $k \geq 1$, a FEM stiffness matrix belongs to $\mathcal{H}(k, P)$.

The reason is that $A|_b = \mathbf{0}$ for all blocks b satisfying (43), since (43) implies that the supports of the basis functions b_i, b_j ($(i, j) \in b$) are disjoint. Remark 9 expresses the fact that \tilde{A} can be considered as \mathcal{H} -matrix. Therefore, we can immediately apply the matrix operations described below. In particular, the inverse matrix can be determined approximately. The latter task requires that A^{-1} has a good approximation $\tilde{B} \in \mathcal{H}(k, P)$. This property is the subject of the next theorem.

Theorem 2. Let $Lu = -\sum_{j=1}^d \partial_{x_j}(c_{j0} \partial_{x_j} u)$ be a uniformly elliptic differential operator whose coefficients are allowed to be extremely nonsmooth: $c_{ij} \in L^\infty(\Omega)$. Let A be a FEM stiffness matrix for this boundary value problem. Then there are approximants $B_k \in \mathcal{H}(k, P)$ so that B_k converges exponentially to A^{-1} (details in Bebendorf and Hackbusch, 2003) (see Chapter 4, this Volume).

4.8 Matrix operations

In the following, we describe the matrix operations that can be performed using \mathcal{H} -matrices. Except for the matrix-vector multiplication, the operations are approximate ones, but the accuracy can be controlled by means of the rank parameter k . For further details and cost estimates, we refer to Grasedyck and Hackbusch (2003).

4.8.1 Matrix-vector multiplication

The matrix-vector product $y \mapsto y + Mx$ is performed by the call $MVM(M, I \times I, x, y)$ of

```

procedure  $MVM(M, b, x, y)$ ;
comment  $b = \tau \times \sigma \in T_{I \times I}$ ,  $M \in \mathbb{R}^{I \times I}$ ,  $x, y \in \mathbb{R}^I$ ;
begin   if  $S_{I \times I}(b) \neq \emptyset$  then for  $b' \in S_{I \times I}(b)$ 
       do  $MVM(M, b', x, y)$ 
       else  $y|_{b'} := y|_{b'} + |M|_b x|_b$ 
end;

```

(48)

The third line of (48) uses the matrix-vector multiplication of an Rk -matrix with a vector (see Remark 6b). The overall arithmetical cost is $O(nk \log n)$.

4.8.2 Matrix-matrix addition, truncation

For $M', M'' \in \mathcal{H}(k, P)$, the exact sum $M := M' + M''$ is obtained by summing $M'|_b + M''|_b$ over all blocks $b \in P$. The problem is, however, that usually $M'|_b + M''|_b$ has rank $2k$, so that a fill-in occurs and M is no longer in the set $\mathcal{H}(k, P)$. Therefore, a truncation of $M|_b = M'|_b + M''|_b$ back to an Rk -matrix $\tilde{M}|_b$ is applied.

Concerning the truncation, we recall the optimal approximation of a general (rectangular) matrix $M \in \mathbb{R}^{n \times m}$ by an Rk -matrix \tilde{M} . Optimality holds with respect to the spectral norm ($\|A\| = \max\{|Ax|/|x| : x \neq 0\} = \sqrt{\lambda_{\max}}$), where λ_{\max} is the maximum eigenvalue of AA^T and the Frobenius norm ($\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$).

Algorithm 1. (a) Calculate the singular-value decomposition $M = U \Sigma V^T$ of M , that is, U, V are unitary matrices, while $\Sigma = \text{diag}(\sigma_1, \dots)$ is a diagonal rectangular matrix containing the singular values $\sigma_1 \geq \sigma_2 \geq \dots$.

(b) Set $\tilde{U} := [U_1, \dots, U_k]$ (first k columns of U), $\tilde{\Sigma} := \text{diag}(\sigma_1, \dots, \sigma_k)$ (first (largest) k singular values), $\tilde{V} := [V_1, \dots, V_k]$ (first k columns of V).

(c) Set $\tilde{A} := \tilde{U} \tilde{\Sigma} \tilde{V}^T \in \mathbb{R}^{n \times m}$ and $\tilde{B} := \tilde{V} \in \mathbb{R}^{m \times k}$ in (46). Then $\tilde{M} = \tilde{A} \tilde{B}^T$ is the best Rk -matrix approximation of M .

We call \tilde{M} a truncation of M to the set of Rk -matrices. The costs are in general $O((\# \tau + \# \sigma)^3)$ operations. In our application, the sum $M := M' + M''$ has rank $K \leq 2k$. Here, we can apply a cheaper singular-value decomposition.

Algorithm 2. Let $M = AB^T$ be an RK -matrix with $A, B \in \mathbb{R}^{n \times K}$ and $K > k$.

(a) Calculate a truncated QR-decomposition $A = Q_A R_A$ of A , that is, $Q_A \in \mathbb{R}^{n \times k}$, $Q_A^T Q_A = I$, and $R_A \in \mathbb{R}^{k \times K}$ upper triangular matrix.

(b) Calculate a truncated QR-decomposition $B = Q_B R_B$ of B , $Q_B \in \mathbb{R}^{n \times k}$, $R_B \in \mathbb{R}^{k \times K}$.

(c) Calculate a singular-value decomposition $U \Sigma V^T$ of the $K \times K$ matrix $R_A R_B^T$.

(d) Set $\tilde{U}, \tilde{\Sigma}, \tilde{V}$ as in Algorithm 1b.

(e) Set $\tilde{A} := Q_A \tilde{U} \tilde{\Sigma} \in \mathbb{R}^{n \times k}$ and $\tilde{B} := Q_B \tilde{V} \in \mathbb{R}^{n \times k}$. Then, $\tilde{M} = \tilde{A} \tilde{B}^T$ is the best Rk -matrix approximation of M .

The truncation from above costs $O(K^2(\# \tau + \# \sigma) + K^3)$ arithmetical operations.

The exact addition $M', M'' \in \mathcal{H}(k, P) \mapsto M := M' + M'' \in \mathcal{H}(2k, P)$ together with the truncation $M \in \mathcal{H}(2k, P) \mapsto \tilde{M} \in \mathcal{H}(k, P)$ is denoted by the *formatted addition*

$$M' \oplus M'' \quad (49)$$

Similarly, the *formatted subtraction* \ominus is defined. The complexity of \oplus and \ominus is $O(nk^2 \log n)$.

4.8.3 Matrix-matrix multiplication

Let $X, Y \in \mathcal{H}(k, P)$. Under the assumption that T_j is a binary tree, both matrices are substructured by $X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$, $Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$, and the product is

$$XY = \begin{bmatrix} X_{11}Y_{11} + X_{12}Y_{21} & X_{11}Y_{12} + X_{12}Y_{22} \\ X_{21}Y_{11} + X_{22}Y_{21} & X_{21}Y_{12} + X_{22}Y_{22} \end{bmatrix}$$

The most costly subproducts are $X_{11}Y_{11}$ and $X_{22}Y_{22}$, since these submatrices have the finest partitioning, whereas $X_{12}, Y_{21}, X_{21}, Y_{12}$ have a coarser format. Performing the products recursively and adding according to Section 4.8.2, we obtain an approximate multiplication $X \odot Y$. Its costs are $O(nk^2 \log^2 n)$ (cf. Hackbusch, 1999). A detailed algorithm can be found in Grasedyck and Hackbusch (2003) and Börm, Grasedyck and Hackbusch (2003).

4.8.4 Inversion

Let $A \in \mathcal{H}(k, P)$. Under the assumption that T_j is a binary tree, we have as above that $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$. The inverse of a 2×2 block matrix can be computed by the block-Gauss elimination (see Hackbusch, 1999), if the principal submatrices are invertible:

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1} A_{12} S^{-1} A_{21} A_{11}^{-1} & -A_{11}^{-1} A_{12} S^{-1} \\ -S^{-1} A_{21} A_{11}^{-1} & S^{-1} \end{bmatrix} \quad (50)$$

with $S = A_{22} - A_{21} A_{11}^{-1} A_{12}$. Applying a recursive procedure Inv , compute $\text{Inv}(A_{11})$ as an approximation of A_{11}^{-1} , invert $\tilde{S} := A_{22} \ominus A_{21} \odot$

$\text{Inv}(A_{11}) \odot A_{12}$ and perform the remaining operations in (50) by means of \otimes and \odot . Again, the precise algorithm is in Börm, Grasedyck and Hackbusch (2003).

The complexity of the computation of the formatted inverse is $O(nk^2 \log^2 n)$ (cf. Hackbusch, 1999; Grasedyck, 2001; Grasedyck and Hackbusch, 2003).

4.9 Examples

4.9.1 BEM case

To demonstrate the advantage of the \mathcal{H} -matrix approach, we consider the simple example of the discretization of the single-layer potential on the unit circle using a Galerkin method with piecewise constant basis functions. The logarithmic kernel function is approximated by the interpolatory approach from Section 2.7.2 (interpolation at Chebyshev points).

The first column of Table 1 contains the number of degrees of freedom ($n = \#I$), the following columns give the relative error $\|A - \tilde{A}\|/\|A\|$ (spectral norm). We observe that the error is bounded independently of the discretization level and that it decreases very quickly when the interpolation order is increased.

The time (SUN Enterprise 6000 using 248 MHz UltraSPARC II) required for matrix-vector multiplications is given in Table 2. We can see that the complexity grows almost linearly in the number of degrees of freedom and rather slowly with respect to the interpolation order.

Finally, we consider the time required for building the \mathcal{H} -matrix representation of the discretized integral operator (see Table 3). The integral of the Lagrange polynomials is computed by using an exact Gauss quadrature formula, while the integral of the kernel function is computed analytically. Once more we observe an almost linear growth of the complexity with respect to the number of degrees of freedom and a slow growth with respect to the interpolation order. Note that even on the specified rather slow processor, the boundary element matrix for more than half a million degrees of freedom can be approximated with an error $< 0.03\%$ in less than half an hour.

Table 1. Approximation error for the single-layer potential.

| n | 1 | 2 | 3 | 4 | 5 |
|-------|------------------|------------------|------------------|------------------|------------------|
| 1024 | 3.57_{10}^{-2} | 2.16_{10}^{-3} | 2.50_{10}^{-4} | 7.88_{10}^{-6} | 2.67_{10}^{-6} |
| 2048 | 3.58_{10}^{-2} | 2.19_{10}^{-3} | 2.51_{10}^{-4} | 7.86_{10}^{-6} | 2.69_{10}^{-6} |
| 4096 | 3.59_{10}^{-2} | 2.20_{10}^{-3} | 2.51_{10}^{-4} | 7.87_{10}^{-6} | 2.68_{10}^{-6} |
| 8192 | 3.59_{10}^{-2} | 2.20_{10}^{-3} | 2.52_{10}^{-4} | 7.76_{10}^{-6} | 2.67_{10}^{-6} |
| 16384 | 3.59_{10}^{-2} | 2.21_{10}^{-3} | 2.53_{10}^{-4} | 7.87_{10}^{-6} | 2.68_{10}^{-6} |

Table 2. Time [s] required for the matrix-vector multiplication (single layer potential).

| n | 1 | 2 | 3 | 4 | 5 |
|--------|-------|-------|-------|-------|-------|
| 1024 | 0.01 | 0.02 | 0.01 | 0.01 | 0.03 |
| 2048 | 0.02 | 0.04 | 0.03 | 0.05 | 0.07 |
| 4096 | 0.05 | 0.11 | 0.09 | 0.12 | 0.17 |
| 8192 | 0.12 | 0.24 | 0.19 | 0.26 | 0.39 |
| 16384 | 0.27 | 0.53 | 0.41 | 0.56 | 0.83 |
| 32768 | 0.57 | 1.15 | 0.90 | 1.23 | 1.90 |
| 65536 | 1.18 | 2.44 | 1.96 | 2.73 | 4.14 |
| 131072 | 2.45 | 5.18 | 4.30 | 5.89 | 8.98 |
| 262144 | 5.15 | 11.32 | 9.14 | 12.95 | 19.78 |
| 524288 | 10.68 | 23.81 | 19.62 | 28.02 | 43.57 |

Table 3. Time [s] required for building the \mathcal{H} -matrix (single layer potential).

| n | 1 | 2 | 3 | 4 | 5 |
|--------|--------|--------|---------|---------|---------|
| 1024 | 0.61 | 0.93 | 1.76 | 3.11 | 5.60 |
| 2048 | 1.25 | 2.03 | 3.85 | 7.04 | 12.94 |
| 4096 | 2.56 | 4.29 | 8.41 | 15.82 | 29.65 |
| 8192 | 5.25 | 9.16 | 18.10 | 35.31 | 66.27 |
| 16384 | 10.75 | 19.30 | 39.32 | 77.47 | 146.65 |
| 32768 | 22.15 | 40.83 | 85.16 | 169.16 | 324.36 |
| 65536 | 45.79 | 87.32 | 185.85 | 368.46 | 702.63 |
| 131072 | 92.64 | 180.73 | 387.63 | 788.06 | 1511.66 |
| 262144 | 189.15 | 378.20 | 854.75 | 1775.85 | 3413.45 |
| 524288 | 388.96 | 795.84 | 1743.66 | 3596.77 | 6950.55 |

4.9.2 FEM case, inverse stiffness matrix

We give a short summary of numerical tests from Grasedyck and Hackbusch (2003) and consider first the Poisson equation $-\Delta u = f$ on the unit square $\Omega = [0, 1]^2$ with zero boundary condition $u = 0$ on $\Gamma = \partial\Omega$. The approximate inverse \tilde{A}^{-1} is computed for different local ranks k . The left part of Table 4 shows the relative error $\|I - A\tilde{A}^{-1}\|$ in the spectral norm for the (formatted) inverse on a uniform grid.

Next we show that the uniformity of the grid and the simple shape of the square do not play any role. The grid from Figure 4 is strongly graded toward the boundary ('boundary concentrated mesh'). For details of the geometrically balanced cluster tree, we refer the reader to Grasedyck and Hackbusch (2003). The complexity of the inversion is reduced compared to that in the uniform case, while the accuracy is enhanced (see right part of Table 4). This resembles the fact that the grid mainly degenerates to a lower dimensional structure (the boundary).

Finally, we give examples showing that the performance is not deteriorated by rough coefficients. Consider the

Table 4. Relative error $\|I - A\tilde{A}^{-1}\|$ in the spectral norm for the (formatted) inverse on a uniform grid (left) and on the boundary concentrated mesh (right).

| k | $n = 4096$ | 16384 | 65536 | 262144 | k | $n = 6664$ | 13568 | 27384 | 55024 | 110312 |
|-----|------------|--------|--------|--------|-----|------------|--------|--------|--------|--------|
| 1 | 2.4 | 8.9 | 2.6+1 | 4.7+1 | 1 | 9.6-2 | 9.9-2 | 7.9-2 | 1.1-1 | 9.4-2 |
| 2 | 5.7-1 | 3.2 | 1.2+1 | 2.7+1 | 2 | 1.3-2 | 1.1-2 | 1.7-2 | 1.9-2 | 1.6-2 |
| 3 | 9.2-2 | 5.2-1 | 2.4 | 1.0+1 | 3 | 3.9-3 | 4.4-3 | 1.7-3 | 4.5-3 | 4.7-3 |
| 4 | 2.0-2 | 9.9-2 | 4.4-1 | 1.91 | 4 | 8.6-5 | 4.7-4 | 1.7-4 | 5.0-4 | 5.1-4 |
| 5 | 2.3-3 | 9.2-3 | 4.0-2 | 1.7-1 | 5 | 8.9-6 | 3.6-5 | 7.6-6 | 4.9-5 | 5.0-5 |
| 6 | 6.4-4 | 3.7-3 | 1.8-2 | 8.4-2 | 6 | 2.1-8 | 9.8-7 | 1.2-6 | 1.3-6 | 1.4-6 |
| 7 | 1.4-4 | 6.9-4 | 2.9-3 | 1.2-2 | 7 | 3.1-10 | 5.0-7 | 1.9-10 | 5.8-7 | 5.9-7 |
| 8 | 7.8-5 | 3.9-4 | 1.8-3 | 7.7-3 | 8 | 1.4-12 | 4.2-10 | 2.1-11 | 2.5-10 | 2.8-10 |
| 9 | 8.5-6 | 4.6-5 | 2.1-4 | 9.4-4 | 9 | 1.0-14 | 2.4-13 | 2.1-14 | 2.7-13 | 2.8-13 |
| 15 | 6.8-9 | 3.3-8 | 1.3-7 | 5.2-7 | | | | | | |
| 20 | 1.7-12 | 1.3-10 | 5.3-10 | 2.5-9 | | | | | | |

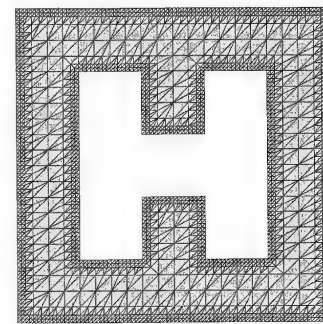


Figure 4. The boundary concentrated mesh.

differential equation

$$-\text{div}(\sigma(x, y)\nabla u(x, y)) = f(x, y) \quad \text{in } \Omega = [0, 1]^2, \\ u = 0 \quad \text{on } \Gamma = \partial\Omega \quad (51)$$

Let $\Omega_1 \subset \Omega$ be the wall-like domain from Figure 5. Let L_σ be the differential operator in (51) with $\sigma(x, y) = \begin{cases} a, & (x, y) \in \Omega_1 \\ 1, & (x, y) \in \Omega \setminus \Omega_1 \end{cases}$. Note that $L_1 = -\Delta$. Table 5 shows the relative accuracy measured for different problem sizes n in the Frobenius norm when approximating the inverse of the respective FEM matrix by an \mathcal{H} -matrix with local rank k . The results demonstrate that the error $\|A^{-1} - \tilde{A}^{-1}\|$ depends on the jump a very weakly.

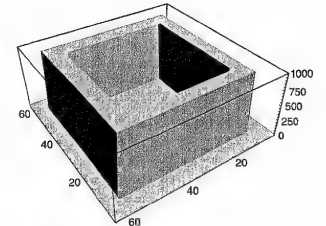


Figure 5. Subdomain Ω_1 of $\Omega = [0, 1]^3$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

4.10 \mathcal{H}^2 -matrices and other variants of \mathcal{H} -matrices

4.10.1 Variable rank, recompression

We may replace the integer k in Definition 8 by a function $k(b)$. Then, the matrix M has to fulfil $\text{rank}(M|_b) \leq k(b)$ for all $b \in P$. A possible nonconstant choice is $k(b) := \alpha * l(b)$, where $l(\cdot)$ is defined by induction: $l(b) = 1$ for leaves $b \in T_{\text{leaf}}$ and $l(b) = 1 + \min\{l(b') : b' \in S_{\text{leaf}}(b)\}$. In this case, $k(b)$ varies between 1 and $\log_2 n$. The low ranks correspond to the (many) small blocks, whereas large ranks occur for the few large blocks. As a result, cost estimates by $O(nk^2 \log^2 n)$ for fixed k may turn into the optimal order $O(n)$ for appropriate variable rank.

Given an \mathcal{H} -matrix M with a certain (variable or constant) local rank, it might happen that the block matrices $M|_b$ can be reduced to lower rank with almost the same accuracy. The standard tool is a singular-value decomposition of $M|_b$. If some of the $k(b)$ singular values $\sigma_1 \geq$

Table 5. Frobenius norm $\|A^{-1} - \tilde{A}^{-1}\|$ of the best approximation to A^{-1} using the local rank k .

| $n = 2304$ | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | $k = 6$ |
|--------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Storage (MB) | 10.2 | 18.9 | 27.6 | 36.2 | | |
| Δ | 4.1 ₁₀ -03 | 5.9 ₁₀ -04 | 1.1 ₁₀ -05 | 1.2 ₁₀ -06 | | |
| L_{10^3} | 6.9 ₁₀ -03 | 9.8 ₁₀ -04 | 1.6 ₁₀ -05 | 2.1 ₁₀ -06 | | |
| L_{10^6} | 6.9 ₁₀ -03 | 9.8 ₁₀ -04 | 1.6 ₁₀ -05 | 1.7 ₁₀ -06 | | |
| $n = 6400$ | | | | | | |
| Storage (MB) | 40.0 | 75.9 | 111.6 | 147.5 | 183.1 | 218.8 |
| Δ | 3.5 ₁₀ -03 | 6.5 ₁₀ -04 | 8.8 ₁₀ -06 | 2.1 ₁₀ -06 | 4.2 ₁₀ -07 | 8.3 ₁₀ -09 |
| L_{10^3} | 5.5 ₁₀ -03 | 1.0 ₁₀ -03 | 1.2 ₁₀ -05 | 3.2 ₁₀ -06 | 5.5 ₁₀ -08 | 1.3 ₁₀ -08 |
| L_{10^6} | 5.6 ₁₀ -03 | 1.0 ₁₀ -03 | 1.2 ₁₀ -05 | 3.1 ₁₀ -07 | 4.7 ₁₀ -08 | 9.1 ₁₀ -09 |
| $n = 14400$ | | | | | | |
| Storage (MB) | 123.4 | 235.7 | 349.6 | 462.0 | 575.9 | 688.2 |
| Δ | 3.2 ₁₀ -03 | 5.9 ₁₀ -04 | 8.9 ₁₀ -06 | 2.3 ₁₀ -06 | 5.5 ₁₀ -08 | 1.5 ₁₀ -08 |
| L_{10^3} | 4.9 ₁₀ -03 | 8.8 ₁₀ -04 | 1.2 ₁₀ -05 | 3.3 ₁₀ -06 | 7.3 ₁₀ -08 | 1.9 ₁₀ -08 |
| L_{10^6} | 5.0 ₁₀ -03 | 8.8 ₁₀ -04 | 1.0 ₁₀ -05 | 3.2 ₁₀ -06 | 6.7 ₁₀ -08 | 9.1 ₁₀ -09 |

$\dots \geq \sigma_{k(b)}$ are sufficiently small, these contributions can be omitted resulting in a smaller local rank $k(b)$.

4.10.2 Uniform \mathcal{H} -matrices

Consider $b = \tau \times \sigma \in T_{\tau, \sigma}$. The submatrix M_b belongs to $\mathbb{R}^{\tau \times \sigma}$. A special subspace of $\mathbb{R}^{\tau \times \sigma}$ is the tensor product space $V_b \otimes W_b = \{vw^T : v \in V_b, w \in W_b\}$ of $V_b \in \mathbb{R}^\tau$ and $W_b \in \mathbb{R}^\sigma$. Note that $T \in V_b \otimes W_b$ implies $\text{rank } T \leq \min(\dim V_b, \dim W_b)$. Hence, we may replace the condition $\text{rank}(M_b) \leq k$ by $M_b \in V_b \otimes W_b$ with spaces V_b, W_b of dimension $\leq k$. The resulting subset of \mathcal{H} -matrices is called the set of uniform \mathcal{H} -matrices.

For the representation of submatrices M_b , one uses corresponding bases $\{v_1, \dots, v_{\dim V_b}\}$ and $\{w_1, \dots, w_{\dim W_b}\}$ of V_b, W_b and defines $V_b := [v_1, \dots, v_{\dim V_b}]$, $W_b := [w_1, \dots, w_{\dim W_b}]$. Then, $M_b = V_b S_b W_b^T$, where the matrix S_b of size $\dim V_b \times \dim W_b$ contains the specific data of M_b that are to be stored.

4.10.3 \mathcal{H}^2 -matrices

The previous class of uniform \mathcal{H} -matrices uses different spaces V_b, W_b for every $b = \tau \times \sigma \in P$. Now, we require that V_b depends only on τ , while W_b depends only on σ . Hence, we may start from a family $V = (V_\tau)_{\tau \in T}$ of spaces $V_\tau \subset \mathbb{R}^\tau$ and require $M_b \in V_\tau \otimes V_\sigma$ for all $b = \tau \times \sigma \in P$. The second, characteristic requirement is the consistency condition

$$V_{\tau'}|_{\tau} \subseteq V_\tau \quad \text{for all } \tau \in T_i \text{ and } \tau' \in S(\tau) \quad (52)$$

that is, $v_{\tau'} \in V_\tau$ for all $v_{\tau'} \in V_{\tau'}$. Let V_τ and $V_{\tau'}$ be the corresponding bases. Owing to (52), there is a matrix $B_{\tau', \tau}$

such that $V_{\tau'}|_{\tau} = V_\tau B_{\tau', \tau}$. Thanks to Definition 5c, V_τ can be obtained from $\{V_{\tau'}, B_{\tau', \tau} : \tau' \in S(\tau)\}$. Hence, the bases V_τ need not be stored, instead the transformation matrices $B_{\tau', \tau}$ are stored. This is an advantage, since their size is $k_\tau \times k_{\tau'}$ with $k_\tau := \dim V_\tau \leq k$ independent of the size of the blocks $b \in P$.

For details on \mathcal{H}^2 -matrices, we refer to Börm and Hackbusch (2002) and Hackbusch, Khoromskij and Sauter (2000). The latter paper considers, for example, the combination of the \mathcal{H}^2 -matrix structure with -variable dimensions k_τ . In Börm, Grasedyck and Hackbusch (2003), the example from Section 4.9.1 is computed also by means of the \mathcal{H}^2 -matrix technique and numbers corresponding to Tables 1 to 3 are given. They show that a slightly reduced accuracy is obtained with considerably less work.

4.11 Applications

We mention three different fields for the application of \mathcal{H} -matrices.

4.11.1 Direct use

In the case of a BEM matrix A , the storage of the n^2 matrix entries must be avoided. Then the approximation of A by an \mathcal{H} -matrix $\tilde{A} \in \mathcal{H}(k, P)$ reduces the storage requirements to almost $\mathcal{O}(n)$. Since in this case, \tilde{A} must carry all information about the BEM problem, the rank k must be chosen high enough (e.g. $k = \mathcal{O}(\log n)$) in order to maintain the accuracy. Second, the matrix-vector multiplication can be performed with almost linear cost.

For the solution of the system of linear equations $Ax = b$, one has two options: (a) Use an iterative scheme that is

based on the matrix-vector multiplication (cg-type methods, multigrid). (b) Compute the approximate inverse \tilde{A}^{-1} (see Section 4.8.4).

The computation of operators like the Steklov operators (Neumann-to-Dirichlet or Dirichlet-to-Neumann map) requires performing the matrix-matrix multiplication \odot from Section 4.8.3.

4.11.2 Rough inverse

In the FEM case, the problem data are given by the sparse stiffness matrix A . The approximate inverse \tilde{A}^{-1} must be accurate enough, if $\tilde{x} := \tilde{A}^{-1}b$ is to be a good approximation of the solution x . However, there is no need to have \tilde{A}^{-1} very accurate. Instead, one can use \tilde{A}^{-1} as 'preconditioner': The iteration

$$x^{m+1} := x^m - \tilde{A}^{-1}(Ax^m - b)$$

can be applied to improve $x^0 := \tilde{A}^{-1}b$. The convergence rate is given by the spectral radius of $I - \tilde{A}^{-1}A$ (cf. Hackbusch, 1994). An upper bound of the spectral radius is the norm $\|I - \tilde{A}^{-1}A\|$ which should be < 1 . For the elliptic example, this norm is given in Table 4.

4.11.3 Matrix-valued problems

There are further problems, where the usual matrix-vector approach is insufficient, since one is interested in matrices instead of vectors. We give some examples that can be solved by means of the \mathcal{H} -matrix technique.

Matrix exponential function

Matrix functions like the matrix exponential can be computed effectively by use of the Dunford-Cauchy representation

$$\exp(A) = \frac{1}{2\pi i} \int_{\Gamma} \exp(z) (zI - A)^{-1} dz \quad (53)$$

where Γ is a curve in the complex plane containing the spectrum of A in its interior. Approximation of the integral by a quadrature rule (z_i : quadrature points) leads to

$$\exp_{\mathcal{H}}(A) = \sum_{i=1}^N e^{-\tau_i z_i} (z_i I - A)^{-1} \quad (54)$$

Since the integration error decreases exponentially with respect to N , one may choose $N = \mathcal{O}(\log^{1/2} 1/\epsilon)$ to obtain an integration error ϵ . The resolvents $(z_i I - A)^{-1}$ are computed due to Section 4.8.4. For further details, we refer to Gavriljuk, Hackbusch and Khoromskij (2002).

Lyapunov equation

There are linear equations for matrices. An example is the Lyapunov equation $AX + XB + C = 0$ for the unknown matrix X , while A, B, C are given. One possible solution uses the representation $X = \int_0^\infty e^{tA} C e^{tB} dt$, provided that the eigenvalues of A, B have negative real parts. Since the dependence of $\exp_{\mathcal{H}}(A)$ on t in (54) is expressed by the scalar factor $e^{-t\lambda}$, one can replace e^{tA}, e^{tB} by $\exp_{\mathcal{H}}(A)$ and $\exp_{\mathcal{H}}(B)$ and perform the integration exactly (cf. Gavriljuk, Hackbusch and Khoromskij, 2002, Section 4.2).

Riccati equation

For optimal control problems, the (nonlinear) Riccati equation

$$A^T X + XA - XFX + G = 0$$

(A, F, G given matrices, X to be determined)

is of interest. In Grasedyck, Hackbusch and Khoromskij (2003), the direct representation of X by means of the matrix-valued sign function is applied. Its iterative computation requires again the inversion, which is provided by the \mathcal{H} -matrix technique.

REFERENCES

- Bebendorf M. Approximation of boundary element matrices. *Numer. Math.* 2000; 86:565-589.
- Bebendorf M and Hackbusch W. Existence of \mathcal{H} -matrix approximations to the inverse FE-matrix of elliptic operators with L^∞ -coefficients. *Numer. Math.* 2003; 95:1-28.
- Bebendorf M and Rjasanov S. Adaptive low-rank approximation of collocation matrices. *Computing* 2003; 70:1-24.
- Börm S and Hackbusch W. \mathcal{H}^2 -matrix approximation of integral operators by interpolation. *Appl. Numer. Math.* 2002; 43:129-143.
- Börm S, Grasedyck L and Hackbusch W. Introduction to hierarchical matrices with applications. *Eng. Anal. Bound. Elem.* 2003; 27:405-422.
- Dahmen W, Prössdorf S and Schneider R. Wavelet approximation methods for pseudodifferential equations II: Matrix compression and fast solution. *Adv. Comput. Math.* 1993; 1:259-335.
- Gavriljuk I, Hackbusch W and Khoromskij BN. \mathcal{H} -matrix approximation for the operator exponential with applications. *Numer. Math.* 2002; 92:83-111.
- Grasedyck L. *Theorie und Anwendungen Hierarchischer Matrizen*. Doctoral thesis, Universität Kiel, 2001.
- Grasedyck L and Hackbusch W. Construction and arithmetics of \mathcal{H} -matrices. *Computing* 2003; 70:295-334.

- Grasedyck L, Hackbusch W and Khoromskij BN. Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices. *Computing* 2003; 70:121–165.
- Greengard L and Rokhlin V. A new version of the fast multipole method for the Laplace equation in three dimensions. *Acta Numer.* 1997; 6:229–269.
- Hackbusch W. *Iterative Solution of Large Sparse Systems*. Springer: New York, 1994 – 2nd German edition: *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. Teubner: Stuttgart, 1993.
- Hackbusch W. *Integral Equations. Theory and Numerical Treatment*. ISNM 128. Birkhäuser: Basel, 1995 – 2nd German edition: *Integralgleichungen. Theorie und Numerik*. Teubner: Stuttgart, 1997.
- Hackbusch W. A sparse matrix arithmetic based on \mathcal{H} -matrices. Part I: Introduction to \mathcal{H} -matrices. *Computing* 1999; 62:89–108.
- Hackbusch W and Khoromskij BN. A sparse matrix arithmetic based on \mathcal{H} -matrices. Part II: Application to multi-dimensional problems. *Computing* 2000; 64:21–47.
- Hackbusch W, Khoromskij BN and Sauter S. On \mathcal{H}^2 -matrices. In *Lectures on Applied Mathematics*, Bungartz H, Hoppe R, Zenger C (eds). Springer: Heidelberg, 2000; 9–29.
- Hackbusch W and Nowak ZP. O złożoności metoda panelej. In *Wydziałowe procesy i systemy*, Marchuk G.I. (ed.), Nauka: Moscow, 1988; 233–244 (conference in Moscow, September 1986).
- Hackbusch W and Nowak ZP. On the fast matrix multiplication in the boundary element method by panel clustering. *Numer. Math.* 1989; 54:463–491.
- Tyrtshnikov E. Mosaico-skeleton approximation. *Calcolo* 1996; 33:47–57.

Chapter 22

Domain Decomposition Methods and Preconditioning

V. G. Korneev¹ and U. Langer²

¹ St. Petersburg State Polytechnical University, St. Petersburg, Russia, and University of Westminster, London, UK

² Johannes Kepler University Linz, Linz, Austria

| | |
|---|-----|
| 1 Introduction | 617 |
| 2 Domain Decomposition History | 619 |
| 3 Fundamentals of Schwarz's Methods | 621 |
| 4 Overlapping Domain Decomposition Methods | 630 |
| 5 Nonoverlapping Domain Decomposition Methods | 633 |
| Acknowledgments | 644 |
| References | 644 |
| Further Reading | 647 |

1 INTRODUCTION

Domain decomposition (DD) methods have been developed for a long time, but most extensively since the first international DD conference held at Paris in 1987. This concerns both the theory and the practical use of DD techniques for creating efficient application software for massively parallel computers. The advances in DD theories and applications are well documented in the proceedings of the annual international DD conferences since 1987 and in numerous papers (see also the DD home page <http://www.ddm.org> for up-to-date information about the annual DD conferences, recent publications, and other DD activities). Two pioneering monographs give an excellent introduction to DD methods from two different points of view: the more algebraic and algorithmic one (Smith, Bjørstad and Gropp,

1996) and the more analytic one (Quarteroni and Valt, 1999). We refer the interested reader also to the survey articles Xu (1992), Le Tallec (1994), Chan and Mathew (1994), and Xu and Zou (1998). We start our chapter with a brief look at the DD history. In this introductory section (Section 2), we provide an exciting journey through the DD history starting in the year 1869 with the classical paper by H.A. Schwarz on the existence of harmonic functions in domains with complicated boundaries, continuing with the variational setting of the alternating Schwarz method by S.L. Sobolev in 1934, looking at the classical finite element (FE) substructuring, or superelement technique intensively used by the engineers in the sixties, and arriving at advanced domain decomposition methods developed mainly during the last 15 years. It is worth mentioning that the classical FE substructuring technique has its roots in calculation methods used in structural mechanics for a long time.

Section 3 gives an introduction to the Schwarz theory (now called Schwarz machinery) that provides a unique framework for constructing and analyzing additive and multiplicative Schwarz methods (preconditioners). Many domain decomposition and multilevel methods (preconditioners) can be put into this framework. Throughout this chapter our model objects are symmetric and positive definite (SPD) systems of algebraic equations typically resulting from finite element discretizations of elliptic problems, such as the heat conduction equation, the potential equation, and the linear elasticity equations. However, the algorithms and some of the results can be extended to more general systems, including systems with indefinite and nonsymmetric system matrices.

In Section 4, overlapping DD methods, which first appeared in Schwarz's original paper in their multiplicative

Section 5 is devoted to nonoverlapping DD methods, certainly most interesting for application. This type of DD methods reflects the classical substructuring finite element technique, where the global sparse finite element system is reduced to a much smaller but denser system (the so-called Schur-complement or interface problem) by condensation of unknowns that are internal for each substructure. Iterative solvers for the interface problems, for example, the conjugate gradient method, are usually most efficient. Together with a good preconditioner they reduce the cost and provide parallelization, in part, by avoiding assembling the Schur complement. Section 5.2 concentrates on various Schur-complement preconditioners. However, the computation of contributions to the Schur complement from the substructures, even without assembling, may, in practice, be more time and memory consuming than direct

solving processes for the original system. For this reason, modern DD algorithms completely avoid the use of Schur complements and use only their preconditioners. Apart from Schur-complement preconditioning, iterative DD methods require efficient solvers for the substructured finite element problems, Dirichlet or others, at each iteration step. They are termed here as *local problems*. Fast direct solvers are hardly available for interesting applications, whereas there exists a variety of well-developed fast iterative solvers, which are easily adapted to local problems in *h*-versions of the finite element method (FEM). They cover many specific situations, for example, subdomains of complicated and specific shapes, orthotropies, and so on. The implementation of local iterative solvers as inexact solvers may be most efficient, but their use for solving the local Dirichlet FE subproblems arising in the so-called extension (prolongation) and restriction operations (from and to the interface, resp.) is very delicate. However, if these procedures are based on bounded discrete extensions (prolongations) and their transposed operations (restrictions), then stability can be proven. This and other topics related to the inexact iterative substructuring are discussed in Section 5.3. We present the (balanced) Neumann–Neumann method as a special Schur-complement preconditioning technique in Section 5.4, and proceed with the finite element tearing and interconnecting (FETI) and Mortar methods in the next two sections. The FETI method requires a conforming triangulation, whereas the FE subspaces are separately given on each substructure including its boundary. The global continuity is then enforced by Lagrange multipliers, resulting in a saddle-point problem.

Finally, let us mention that this contribution cannot cover all aspects of domain decomposition techniques and their applications. For instance, the p - and the hp -versions of the FEM have some specific features that are not discussed in this contribution in detail; see **Chapter 5, Chapter 6 of this Volume** and **Chapter 3, Volume 3** for more information on these topics. Other discretization techniques like the boundary element methods (BEM) are also not discussed in this paper (see **Chapter 12** and **Chapter 21 of this Volume**). The coupling of FEM and BEM is naturally based on DD techniques (see **Chapter 13, this Volume**). We also refer to the corresponding publications, which have mostly appeared quite recently. The field of the application of DD methods is now very wide. Here, we especially refer to the proceedings of the annual DD conferences mentioned above.

Schwarz (1869) investigated the existence of harmonic functions in domains Ω with complicated boundaries $\partial\Omega$. Given some boundary function g , find a function u such that

$$u(x) = g(x) \quad \forall x \in \partial\Omega \quad (2)$$

$$u(x) = 0 \quad \forall x \in \partial\Omega \quad (4)$$

```

 $u^0 \in C^2(\Omega) \cap C(\overline{\Omega})$  given initial guess:  $u^0 = 0$  on  $\partial\Omega$  (initialization loop)
for  $n = 0$  step 1 until Convergence do (begin iteration loop)
  First step: (update in  $\Omega_1$ )
  Define  $\tilde{u}^{n+1/2} \in C^2(\Omega_1) \cap C(\overline{\Omega}_1)$ :  $-\Delta \tilde{u}^{n+1/2} = f$  in  $\Omega_1$ ,  $\tilde{u}^{n+1/2} = u^n$  on  $\partial\Omega_1$ ,
   $u^{n+1/2}(x) = \tilde{u}^{n+1/2}(x) \quad \forall x \in \overline{\Omega}_1$ ,
   $u^{n+1/2}(x) = u^n(x) \quad \forall x \in \overline{\Omega}_2 \setminus \overline{\Omega}_1$ .
  Second step: (update in  $\Omega_2$ )
  Define  $\tilde{u}^{n+1} \in C^2(\Omega_2) \cap C(\overline{\Omega}_2)$ :  $-\Delta \tilde{u}^{n+1} = f$  in  $\Omega_2$ ,  $\tilde{u}^{n+1} = u^{n+1/2}$  on  $\partial\Omega_2$ ,
   $u^{n+1}(x) = \tilde{u}^{n+1}(x) \quad \forall x \in \overline{\Omega}_2$ ,
   $u^{n+1}(x) = u^{n+1/2}(x) \quad \forall x \in \overline{\Omega}_1 \setminus \overline{\Omega}_2$ .
end for (end iteration loop)

```

end for (end iteration loop)

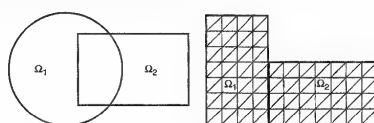


Figure 1. Overlapping and nonoverlapping DD samples.

Algorithm 1 now describes the Alternating Schwarz Method for solving the boundary value problem (3)–(4) as a model problem. The convergence analysis was done by Schwarz (1869) using the maximum principle (see also Nevanlinna, 1939).

Sobolev (1936) gave the variational setting of the Alternating Schwarz Method for solving linear elasticity problems in the form of an alternating minimization procedure in Ω_1 and Ω_2 . Algorithm 2 provides the corresponding variational formulation that is nothing else but the weak formulation of the Alternating Schwarz Algorithm 1. Sobolev (1936) proved convergence of the Variational Alternating Schwarz Algorithm 2 in L_2 to the weak solution of the boundary value problem (3)–(4) provided by its variational formulation: Find $u \in V_0 = H_0^1(\Omega)$ such that

$$a(u, v) = (f, v) \quad \forall v \in V_0 \quad (7)$$

where the bilinear form $a(\cdot, \cdot): V_0 \times V_0 \rightarrow \mathbb{R}$ and the linear form $(f, \cdot): V_0 \rightarrow \mathbb{R}$ are defined as follows:

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx$$

$$(f, v) = \int_{\Omega} f(x)v(x) \, dx \quad (8)$$

Mikhlin (1951) showed uniform convergence in every closed subdomain of the domain Ω . Since then, the Schwarz Alternating Method was studied by many authors (see, for example, Morgenstern, 1956 and Babuska, 1958).

As just mentioned, the solution of the variational equations (5) and (6) in Algorithm 2 is equivalent to the alternating minimization of the energy functional $E(\cdot) = (1/2)a(\cdot, \cdot) - (f, \cdot)$ in V_1 and V_2 respectively, that is,

$$E(u^{n+1/2}) = \min_{w \in V_1} E(u^n + w)$$

$$E(u^{n+1}) = \min_{w \in V_2} E(u^{n+1/2} + w) \quad (9)$$

Moreover, if we introduce the orthoprojections $P_i: V \rightarrow V_i$ by the identities

$$a(P_i u, v) = a(u, v) \quad \forall v \in V_i, \quad \forall u \in V \quad (10)$$

then we immediately observe from (5) and (6) that $u^{n+1/2} = P_1(u - u^n)$ and $u^{n+1} = P_2(u - u^{n+1/2})$ respectively. Thus, the iteration error $z^n = u - u^n$ satisfies the recurrence relation

$$z^{n+1} = (I - P_2)(I - P_1)z^n = (I - P_1 - P_2 + P_2 P_1)z^n \quad (11)$$

that is nothing but an alternating orthoprojection of the iteration error to V_1^{\perp} and V_2^{\perp} . This alternating projection procedure can obviously be generalized to a decomposition of V into many subspaces, to finite element subspaces, to other problems, and so on. Owing to the multiplicative nature of the error transition, these kind of alternating projection procedures are nowadays called multiplicative Schwarz methods (MSM). The main drawback of the MSM is connected with the sequential character of this procedure that makes the parallelization difficult. To overcome this drawback, additive versions of Schwarz algorithms were proposed. These observations led to the modern theory of Schwarz methods that has been developed during the last 15 years (see Section 3).

The substructuring technique developed by mechanical engineers for the finite element analysis of complex structures in the sixties (see e.g. Przemieniecki, 1963) is usually recognized for the other main root of the modern numerical DD algorithms. In the classical finite element substructuring technique, the computational domain Ω is decomposed into J nonoverlapping subdomains (substructures) Ω_j ($j = 1, 2, \dots, J$) such that $\Omega = \bigcup_{j=1}^J \bar{\Omega}_j$ and $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$, and each subdomain Ω_i is divided into finite elements δ_i such that this discretization process results in a conform triangulation of Ω . In the following, the indices 'C' and 'I' correspond to the nodes belonging to the coupling boundaries (interfaces, skeleton) $\Gamma_C = \bigcup_{j=1}^J \partial\Omega_j \setminus \Gamma_D$ and to the interior $\Omega_I = \bigcup_{j=1}^J \Omega_j$ of the subdomains, respectively, where Γ_D is that part of $\partial\Omega$ where Dirichlet-type boundary conditions are given (see

also the right part of Figure 1). Boundaries with natural (Neumann, Robin) boundary conditions will be handled as coupling boundaries.

Let us define the usual FE nodal basis

$$\Phi = [\Phi_C, \Phi_I]$$

$$= [\phi_1, \dots, \phi_{N_C}, \phi_{N_C+1}, \dots, \phi_{N_C+N_I}, \dots, \phi_{N_C+N_I+N_I}] \quad (12)$$

where the first N_C basis functions belong to Γ_C , the next N_{I_1} to Ω_1 , the next N_{I_2} to Ω_2 , and so on, with $N_I = N_{I_1} + N_{I_2} + \dots + N_{I_J}$. The FE space $V = V_h$ is obviously a finite-dimensional subspace of the variational function space V_0 . Once the FE basis Φ is chosen, the FE scheme leads to a large-scale sparse system $Ku = f$ of finite element equations with the SPD stiffness matrix K provided that the bilinear form has the corresponding properties. Owing to the arrangement of the basis functions made above, the FE system can be rewritten in the block form

$$\begin{pmatrix} K_C & K_{CI} \\ K_{IC} & K_I \end{pmatrix} \begin{pmatrix} u_C \\ u_I \end{pmatrix} = \begin{pmatrix} f_C \\ f_I \end{pmatrix} \quad (13)$$

where $K_I = \text{diag}(K_{I_j})_{j=1,2,\dots,J}$ is block diagonal. The block diagonal entries K_{I_j} are of the dimension $N_{I_j} \times N_{I_j}$ and arise from the FE approximation to the PDE considered in Ω_j under homogeneous Dirichlet boundary conditions on $\partial\Omega_j$. Owing to the block diagonal structure of K_I , one can eliminate the internal subdomain unknowns u_I in parallel by block Gaussian elimination. Solving the resulting Schur-complement problem, we obtain the coupling node (interface) unknowns that allow us to define the internal subdomain unknowns finally. This classical FE substructuring algorithm is described in detail by Algorithm 3. The classical FE substructuring Algorithm 3 is well suited for parallel implementation, but very expensive with respect to the arithmetical operations, especially the forming of the Schur-complement matrix S_C is very time consuming. If an iterative solver is used for solving the Schur-complement problem, the forming of the Schur-complement matrix can be avoided because only the matrix-by-vector operation

Algorithm 3. Classical substructuring algorithm for solving (13).

$$\begin{aligned} \tilde{u}_I &= K_I^{-1} f_I, \text{ that is, solve } K_I \tilde{u}_I = f_I && \text{(elimination of the internal unknowns in parallel)} \\ g_C &= f_C - K_{CI} \tilde{u}_I && \text{(forming of the right-hand side)} \\ S_C &= K_C - K_{CI} K_I^{-1} K_{IC} && \text{(forming of the Schur complement)} \\ u_C &= S_C^{-1} g_C && \text{(solving the Schur-complement problem)} \\ u_I &= \tilde{u}_I - K_I^{-1} K_{IC} u_C && \text{(determination of the internal unknowns in parallel)} \end{aligned}$$

$S_C \cdot u_C^*$ is required. This leads us to the iterative substructuring methods, which are the starting point for the modern nonoverlapping domain decomposition methods discussed in Section 5. Iterative nonoverlapping DD algorithms and their mathematical studies appeared in the seventies. The first consistent analysis of the role of Poincaré-Steklov operators (the operator analogue of the Schur-complement matrix S_C) in such algorithms was presented in the book by Lebedev and Agoshkov (1983).

3 FUNDAMENTALS OF SCHWARZ'S METHODS

3.1 Preliminaries

Let us consider a symmetric, elliptic (coercive), and bounded (continuous) abstract variational problem of the following form: Given $f \in V_0'$, find $u \in V_0$ such that the variational equation

$$a(u, v) = (f, v) \quad \forall v \in V_0 \quad (14)$$

holds for all test functions from some Hilbert space V_0 equipped with the scalar product $(\cdot, \cdot)_{V_0}$ and the corresponding norm $\|\cdot\|_{V_0}$. The bilinear form $a(\cdot, \cdot): V_0 \times V_0 \rightarrow \mathbb{R}$ is supposed to be symmetric, that is,

$$a(u, v) = a(v, u) \quad \forall u, v \in V_0 \quad (15)$$

V_0 -elliptic (V_0 -coercive), that is, there exists some positive constant μ_1 such that

$$\mu_1 \|v\|_{V_0}^2 \leq a(v, v) \quad \forall v \in V_0 \quad (16)$$

and V_0 -bounded (V_0 -continuous), that is, there exists some positive constant μ_2 such that

$$a(u, v) \leq \mu_2 \|u\|_{V_0} \|v\|_{V_0} \quad \forall u, v \in V_0 \quad (17)$$

The value of the bounded (continuous) linear functional f from the dual space V_0^* at some $v \in V_0$ is denoted by $\langle f, v \rangle$. Sometimes $\langle \cdot, \cdot \rangle : V_0^* \times V_0 \rightarrow \mathbb{R}$ is called duality product. Owing to Lax–Milgram's lemma, the V_0 -ellipticity (16) and the V_0 -boundness (17) ensure the existence and uniqueness of the solution of the abstract variational problem (14), (see e.g. Ciarlet, 1978).

Abstract variational formulations of the form (14) cover a lot of practically very important formally self-adjoint, elliptic boundary value problems. The Dirichlet boundary value problem for the Poisson equation introduced in Section 2 is certainly the most prominent representative of this class. Other representatives are the stationary heat conduction equation (Example 1), the linear elasticity problem (Example 2), linearized mechanical problems, and linear magneto- and electrostatic boundary value problems.

Example 1 Stationary heat conduction problem. Given some heat source intensity function $f \in L_2(\Omega)$, find the temperature field $u \in V_0 = H_0^1(\Omega)$ such that the variational equation (15) holds with the bilinear form $a(\cdot, \cdot) : V_0 \times V_0 \rightarrow \mathbb{R}$ and the linear form $\langle f, \cdot \rangle : V_0 \rightarrow \mathbb{R}$ defined by the identities

$$\begin{aligned} a(u, v) &= \int_{\Omega} \alpha(x) \nabla u(x) \cdot \nabla v(x) \, dx \\ \langle f, v \rangle &= \int_{\Omega} f(x) v(x) \, dx \end{aligned} \quad (18)$$

respectively. As everywhere else in this chapter, if it is not defined otherwise, the computational domain $\Omega \in \mathbb{R}^d$ ($d = 1, 2, 3$) is assumed to be bounded and sufficiently smooth (e.g. with a Lipschitz boundary). The given heat conduction coefficient $\alpha(\cdot)$ is supposed to be uniformly positive and bounded. The symmetry of the bilinear form is obvious. The V_0 -ellipticity (16) and the V_0 -boundness (17) directly follow from the Friedrichs and Cauchy inequalities, respectively (see e.g. Ciarlet, 1978). Here, we consider only homogeneous Dirichlet boundary conditions (vanishing temperature u on the boundary $\partial\Omega$ of Ω). Other boundary conditions (Neumann, Robin, mixed) can be treated in the same way. In many practical cases, the coefficient of conductivity has significant jumps, so that the ratio μ_2/μ_1 is very large. This requires DD algorithms which are robust with respect to μ_2/μ_1 (see Section 5 for robust DD methods). We mention that the heat conduction equation formally describes many other stationary processes, like diffusion or filtration in porous media with variable permeability.

Example 2 The static linear elasticity problem. Given volume forces $f = (f_1, \dots, f_d)^T$ in Ω and surface tractions $t = (t_1, \dots, t_d)^T$ on some part $\Gamma_N = \partial\Omega \setminus \Gamma_D$ of the boundary $\partial\Omega$, find the displacement $u = (u_1, \dots, u_d)^T \in V_0 = \{v = (v_1, \dots, v_d)^T : v_i \in H^1(\Omega), v_i = 0 \text{ on } \Gamma_D, i = 1, \dots, d\}$ of the elastic body $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) clamped at Γ_D such that the variational equation (15) holds with the bilinear form $a(\cdot, \cdot) : V_0 \times V_0 \rightarrow \mathbb{R}$ and the linear form $\langle f, \cdot \rangle : V_0 \rightarrow \mathbb{R}$ defined by the identities

$$\begin{aligned} a(u, v) &= \int_{\Omega} \sum_{i,j,k,l=1}^d \varepsilon_{ij}(x) D_{ijkl}(x) \varepsilon_{kl}(x) \, dx \\ \langle f, v \rangle &= \int_{\Omega} \sum_{i=1}^d f_i(x) v_i(x) \, dx + \int_{\Gamma_N} \sum_{i=1}^d t_i(x) v_i(x) \, ds \end{aligned} \quad (19)$$

and

$$(20)$$

respectively, where the $\varepsilon_{ij} = (1/2)(\partial u_i/\partial x_j + \partial u_j/\partial x_i)$ denote the linearized strains. The given matrix $D(x) = (D_{ijkl}(x))$ of the elastic coefficients is supposed to be symmetric, uniformly positive definite, and uniformly bounded. These assumptions together with Korn's, Friedrichs', and Cauchy's inequalities ensure the symmetry (15), V_0 -ellipticity (16) and V_0 -boundness (17) of the bilinear form. If the volume forces and surface tractions are chosen in such a way that the corresponding linear functional (20) is continuous on V_0 , then the Lax–Milgram lemma again provides existence and uniqueness of the solution of the static linear elasticity problem (see e.g. Ciarlet, 1978 and Korneev and Langer, 1984).

We now approximate the abstract variational equation (14) by some FE Galerkin scheme. Let V_h be some finite-dimensional FE subspace of the space V_0 spanned by some basis $\Phi_h := \{\phi_1, \phi_2, \dots, \phi_{N_h}\}$, that is, $V_h = \text{span} \Phi_h \subset V_0$, where h denotes some usual discretization parameter such that the number of unknowns N_h behaves like $O(h^{-d})$ as h tends to 0. Here and in the following, we assume that the FE discretization is based on some quasiuniform triangulation. Note that we use Φ_h as symbol for the set of basis functions $\{\phi_i\}_{i=1, \dots, N_h}$ as well as for the FE-Galerkin isomorphism ($u_h \leftrightarrow u_h$)

$$u_h = \Phi_h u_h := \sum_{i=1}^{N_h} u_i \phi_i \quad (21)$$

mapping some vector of nodal parameters $u_h = (u_i)_{i=1, \dots, N_h} \in \mathbb{R}^{N_h}$ to the corresponding FE function $u_h \in V_h$. Now the FE-Galerkin solution of the variational equation (14) is nothing but the solution of (14) on the FE subspace V_h :

Given $f \in V_0^*$, find $u_h \in V_h$ such that

$$a(u_h, v_h) = \langle f, v_h \rangle \quad \forall v_h \in V_h \quad (22)$$

Once the basis Φ_h is chosen, the FE scheme (22) is equivalent to the following system of FE equations: Find the nodal parameter vector $u_h \in \mathbb{R}^{N_h}$ corresponding to the FE solution u_h by the Galerkin isomorphism (21) as the solution of the system

$$K_h u_h = f_h \quad (23)$$

where the stiffness matrix K_h and the load vector f_h are generated from the identities

$$\begin{aligned} (K_h u_h, v_h) &= a(\Phi_h u_h, \Phi_h v_h) \\ &= a(u_h, v_h) \quad \forall u_h, v_h \leftrightarrow u_h, v_h \in \mathbb{R}^{N_h} \end{aligned} \quad (24)$$

and

$$(f_h, v_h) = \langle f, v_h \rangle \quad \forall v_h \leftrightarrow v_h \in \mathbb{R}^{N_h} \quad (25)$$

respectively. Here $(f_h, v_h) = (f_h, v_h)_{\mathbb{R}^{N_h}} = f_h^T v_h$ denotes the Euclidean scalar product in \mathbb{R}^{N_h} .

In order to simplify the notation, we skip the subscript h in the following. Since we are primarily interested in the solution of the FE equations, there will be no confusion of some FE function $u = u_h \in V = V_h$ and functions u from the space V_0 . Let us now assume that the FE space

$$V = V_h = \sum_{j=1}^J V_j \quad (26)$$

can be split into a (not necessarily direct) sum of the J subspaces

$$V_j = \text{span} \Psi_j = \text{span} \Phi_j V_j, \quad j = 1, 2, \dots, J \quad (27)$$

where the basis $\Psi_j = \{\psi_{j1}, \psi_{j2}, \dots, \psi_{jN_j}\} = \Phi_j V_j$ of the subspace V_j is obtained from the original basis Φ by the $N \times N_j$ basis transformation matrix V_j . Therefore, $N_j = \dim V_j = \text{rank } V_j$ and $\sum_{j=1}^J N_j \geq N$.

The orthoprojection $P_j : V \rightarrow V_j$ of the space V onto its subspace V_j with respect to the energy inner product $a(\cdot, \cdot)$ is uniquely defined by the identity

$$a(P_j u, v_j) = a(u, v_j) \quad \forall v_j \in V_j, \quad \forall u \in V \quad (28)$$

Sometimes the orthoprojection P_j is called the Ritz, or energy projection. As orthoprojection, $P_j = P_j^*$ is self-adjoint with respect to the energy inner product, that is,

$$a(P_j u, v) = a(u, P_j v) \quad \forall u, v \in V \quad (29)$$

and satisfies the projection relation $P_j^2 = P_j$. It is easy to see from (24) and (27) that the orthoprojection $P_j u$ of some $u \leftrightarrow u$ can be computed by the formula

$$P_j u = \Phi V_j u_j = \Phi V_j (V_j^T K V_j)^{-1} V_j^T K u \quad (30)$$

that is, u_j is obtained from the solution of a smaller system with the $N_j \times N_j$ system matrix $V_j^T K V_j$ and the right-hand side $V_j^T K u$. Similarly, replacing the original energy inner product $a(\cdot, \cdot)$ on the left-hand side of the identity (28) by some individual symmetric, V_j -elliptic and V_j -bounded bilinear form $a_j(\cdot, \cdot) : V_j \times V_j \rightarrow \mathbb{R}$, we define some projection-like operator $\tilde{P}_j : V \rightarrow V_j$ that is self-adjoint but in general $\tilde{P}_j^2 \neq \tilde{P}_j$. Thus, \tilde{P}_j is not an orthoprojection. Again, $\tilde{P}_j u$ can be easily calculated by the formula

$$\tilde{P}_j u = \Phi V_j u_j = \Phi V_j C_j^{-1} V_j^T K u \quad (31)$$

where the $N_j \times N_j$ matrix C_j is generated from the identity

$$(C_j u_j, v_j) = a_j(u_j, v_j) \quad \forall u_j, v_j \leftrightarrow u_j, v_j \in \mathbb{R}^{N_j} \quad (32)$$

in the same way as K was generated above from the original bilinear form $a(\cdot, \cdot)$.

3.2 Schwarz methods and preconditioners for elliptic variational problems

In the next three sections, we describe various Schwarz algorithms and the corresponding preconditioners. We mention that the Schwarz algorithms are completely defined by the space splitting (26), the subspace bilinear forms $a_j(\cdot, \cdot)$, and the arrangement of the projection-like operations. Finally, we present some interesting examples.

3.2.1 Additive algorithms

The (inexact) Additive Schwarz Method (ASM) corresponding to the space splitting (26) and to the subspace bilinear forms $a_j(\cdot, \cdot)$ can be written in the form of an iteration process in the FE space V (function version) and in \mathbb{R}^N (vector/matrix version) as shown in Algorithm 4. Replacing $a_j(\cdot, \cdot)$ by $a(\cdot, \cdot)$, C_j by $V_j^T K V_j$, and \tilde{P}_j by P_j in Algorithm 4, we arrive at the so-called exact ASM. In this sense, we consider the exact ASM as special case of the inexact ASM presented in Algorithm 4.

The iteration error $z^n = u - u^n \in V$ now satisfies the error iteration scheme

$$z^{n+1} = E z^n \quad (44)$$

resulting in the energy norm iteration error estimate

$$\|z^{n+1}\|_a \leq \|E\|_a \|z^n\|_a \quad (45)$$

with the MSM error propagation operator (iteration operator)

$$E = (I - \tilde{P}_J)(I - \tilde{P}_{J-1}) \cdots (I - \tilde{P}_1) \quad (46)$$

Therefore, the convergence rate of the MSM iteration is completely defined by the (operator) energy norm $\|E\|_a$ of the MSM error propagation operator E . The same is true for the vector version with respect to the error propagation matrix (iteration matrix) $\mathbf{E} = (\mathbf{I} - \mathbf{V}_J \mathbf{C}_J^{-1} \mathbf{V}_J^T \mathbf{K}) \cdots (\mathbf{I} - \mathbf{V}_1 \mathbf{C}_1^{-1} \mathbf{V}_1^T \mathbf{K})$. Convergence rate estimates of the form

$$\|E\|_a = \|\mathbf{E}\|_K \leq \rho_{\text{MSM}} < 1 \quad (47)$$

will be presented in Section 3.3. Unfortunately, the MSM preconditioner

$$\mathbf{C} = \mathbf{K}(\mathbf{I} - \mathbf{E})^{-1} \quad (48)$$

is not symmetric and, therefore, cannot be used in the PCG as a preconditioner.

However, repeating the subspace corrections in Algorithm 5 in the reverse direction, we arrive at the so-called symmetric Multiplicative Schwarz Method (sMSM) that is characterized by the error propagation operator

$$E = (I - \tilde{P}_1) \cdots (I - \tilde{P}_{J-1})(I - \tilde{P}_J)(I - \tilde{P}_{J-1}) \cdots (I - \tilde{P}_1) \quad (49)$$

resp. the error propagation matrix

$$\mathbf{E} = (\mathbf{I} - \mathbf{V}_1 \mathbf{C}_1^{-1} \mathbf{V}_1^T \mathbf{K}) \cdots (\mathbf{I} - \mathbf{V}_J \mathbf{C}_J^{-1} \mathbf{V}_J^T \mathbf{K}) \times (\mathbf{I} - \mathbf{V}_J \mathbf{C}_J^{-1} \mathbf{V}_J^T \mathbf{K}) \cdots (\mathbf{I} - \mathbf{V}_1 \mathbf{C}_1^{-1} \mathbf{V}_1^T \mathbf{K}) \quad (50)$$

resulting in a symmetric preconditioner (48). Mention that, in the exact version of the sMSM, the subspace correction in V_j has to be carried out only once because $(I - P_j)(I - P_j) = (I - P_j)$.

3.2.3 Hybrid algorithms

There are a lot of useful hybrid Schwarz algorithms corresponding to various possibilities of the arrangement of the subspace correction in an additive and multiplicative

manner. The algorithm can be completely defined by its iteration operator E resp. the iteration matrix \mathbf{E} .

For instance, the iteration operator

$$E = (I - \tilde{P}_1)(I - \tau(\tilde{P}_2 + \cdots + \tilde{P}_J))(I - \tilde{P}_1) \quad (51)$$

corresponds to the Hybrid Schwarz Method described by Algorithm 6 (function version only). The error propagation operator (51) resp. Algorithm 6 correspond to a symmetric preconditioner of the form (48). Note that the error propagation operator (51) as well as the corresponding iteration matrix are self-adjoint with respect to the corresponding energy inner products.

3.2.4 Examples

We recall that the Schwarz algorithm (preconditioner) is uniquely defined by the space splitting $V = \sum_{j=1}^J V_j$, the subspace bilinear forms $a_j(\cdot, \cdot)$ and the arrangement of the projection-like operations.

Example 3 Nodal basis splitting. For $j = \overline{1, J}$ and $J = N$, we define the one-dimensional ($N_j = \dim V_j = 1$) subspaces

$$V_j = \text{span}\{\phi_j\} = \text{span}\Phi V_j \quad (52)$$

of V giving the so-called nodal basis splitting $V = \sum_{j=1}^N V_j$, where $V_j = e_j$ is the j th unit vector $e_j = (0, \dots, 0, 1, 0, \dots, 0)^T$. Therefore,

$$\mathbf{V}_j^T \mathbf{K} \mathbf{V}_j = e_j^T \mathbf{K} e_j = K_{jj} \quad (53)$$

Taking this relation into account, we observe that the ASM preconditioner (36) in its exact version ($\mathbf{C}_j = \mathbf{V}_j^T \mathbf{K} \mathbf{V}_j$) is nothing but the well-known Jacobi preconditioner (diagonal scaling)

$$\mathbf{C}^{-1} = \sum_{j=1}^J e_j K_{jj}^{-1} e_j^T = \mathbf{D}^{-1} = (\text{diag}(\mathbf{K}))^{-1} \quad (54)$$

and the exact ASM coincides with the classical (damped) Jacobi method. Furthermore, the exact MSM corresponding to our nodal basis splitting gives us the classical Gauss-Seidel method. Indeed, from the exact version of the MSM Algorithm 5 and from (53), we see that the j th subspace correction step in the n th iteration step

$$\begin{aligned} u_j^{n+(j/J)} &= u^{n+(j-1)/J} + \mathbf{V}_j w_j^{n+(j/J)} \\ &= u^{n+(j-1)/J} + e_j (e_j^T \mathbf{K} e_j)^{-1} e_j^T (f - \mathbf{K} u^{n+(j-1)/J}) \end{aligned} \quad (55)$$

Algorithm 6. Hybrid Schwarz method corresponding to the error propagation operator (51).

```

 $u^0 = \Phi u^0 \in V$  given initial guess and  $\tau$  given iteration parameter      {initialization}
                                                                              {begin iteration loop}

for  $n = 0$  step 1 until Convergence do                                {first multiplicative subspace correction in  $V_1$ }
     $w_1^{n,1} \in V_1 : a_1(w_1^{n,1}, v_1) = (f, v_1) - a(u^n, v_1) \quad \forall v_1 \in V_1$ 
     $u^{n,1} = u^n + w_1^{n,1} = u^n + \tilde{P}_1(u - u^n)$                                 {additive subspace corrections in remaining subspaces}

    for all  $j \in \{2, \dots, J\}$  in parallel do
         $w_j^{n,1} \in V_j : a_j(w_j^{n,1}, v_j) = (f, v_j) - a(u^{n,1}, v_j) \quad \forall v_j \in V_j$ 
    end for
     $u^{n,1} = u^{n,1} + \tau \sum_{j=2}^J w_j^{n,1} = u^{n,1} + \tau \sum_{j=2}^J \tilde{P}_j(u - u^{n,1})$ 
                                                                              {second multiplicative subspace correction in  $V_1$ }

     $w_1^{n,2} \in V_1 : a_1(w_1^{n,2}, v_1) = (f, v_1) - a(u^{n,2}, v_1) \quad \forall v_1 \in V_1$ 
     $u^{n+1} = u^{n,1} + w_1^{n,2} = u^{n,2} + \tilde{P}_1(u - u^{n,2})$ 

end for                                                                    {end iteration loop}

```

updates only the j th component of the iterate

$$\begin{aligned} u_j^{n+(j/J)} &= \frac{1}{K_{jj}} \left(f_j - \sum_{i=1}^{j-1} K_{ji} u_i^{n+(j-1)/J} \right. \\ &\quad \left. - \sum_{i=j+1}^J K_{ji} u_i^{n+(j-1)/J} \right) \quad j = 1, 2, \dots, J \quad (J = N) \end{aligned} \quad (56)$$

Formulas (56) exactly describe the Gauss-Seidel iteration procedure. In this sense, we look at the ASM and the MSM as the natural generalizations of the Jacobi method and the Gauss-Seidel method, respectively. In an analogous way, the symmetric Gauss-Seidel method corresponds to the exact sMSM. Similar to the SOR and the SSOR methods, we can also introduce overrelaxation parameters into MSM and sMSM aiming at the improvement of convergence (see Griebel and Oswald (1995) for related results).

Therefore, the Jacobi iteration and the Gauss-Seidel method are the classical prototypes of the ASM and the MSM, respectively. Further examples are given in Section 4.3 and correspond to multilevel splittings of the finite element space V . The most prominent one is the so-called BPX preconditioner.

3.3 Spectral equivalence estimates and convergence analysis

In this section, we present some convergence results for the Schwarz methods and some spectral equivalence results for the corresponding Schwarz preconditioners. The main condition for creating good Schwarz methods (preconditioners) consists in a stable splitting of the (FE) space V into subspaces (V_j). We first consider the simple case of splitting V into a direct sum of two subspaces V_1 and V_2 . This case is very important for the nonoverlapping domain decomposition methods studied in Section 5. Finally, we give some result for the general case of splitting V into J subspaces.

3.3.1 The simple case: splitting into a direct sum of two subspaces

Let us consider the simple case of splitting

$$V = V_1 + V_2 \quad \text{and} \quad V_1 \cap V_2 = \{0\} \quad (57)$$

into the two (nontrivial) subspaces $V_1 = \text{span}\Phi V_1$ and $V_2 = \text{span}\Phi V_2$, and let us define the cosine of the angle

between V_1 and V_2 :

$$\gamma = \cos \angle(V_1, V_2) := \sup_{v_1 \in V_1 \setminus \{0\}, v_2 \in V_2 \setminus \{0\}} \frac{a(v_1, v_2)}{\|v_1\|_a \|v_2\|_a} < 1 \quad (58)$$

corresponding to the sharp constant γ in the so-called strengthened Cauchy inequality:

$$|a(v_1, v_2)| \leq \gamma \|v_1\|_a \|v_2\|_a \quad \forall v_1 \in V_1, \forall v_2 \in V_2 \quad (59)$$

The splitting (57) is called stable if and only if the constant γ stays less than 1 for growing dimensions $N = N_h \rightarrow \infty$ ($h \rightarrow 0$).

The following lemma gives some useful relations for computing or estimating γ .

Lemma 1. *The following relations are valid:*

$$\cos \angle(V_1, V_2) = \cos \angle(V_1^\perp, V_2^\perp) \quad (60)$$

$$\sup_{v_1 \in V_1 \setminus \{0\}, v_2 \in V_2 \setminus \{0\}} \frac{a(v_1, v_2)}{\|v_1\|_a \|v_2\|_a} = \sup_{v_1 \in V_1 \setminus \{0\}, v_2 \in V_2 \setminus \{0\}} \frac{2a(v_1, v_2)}{\|v_1\|_a^2 + \|v_2\|_a^2} \quad (61)$$

$$\sup_{v_1 \in V_1 \setminus \{0\}, v_2 \in V_2 \setminus \{0\}} \frac{a(v_1, v_2)}{\|v_1\|_a \|v_2\|_a} = \sup_{v_1 \in \mathbb{R}^{n_1} \setminus \{0\}, v_2 \in \mathbb{R}^{n_2} \setminus \{0\}} \frac{((V_1^T K V_2)(V_2^T K V_2)^{-1}(V_2^T K V_1)v_1, v_2)}{(V_1^T K V_1 v_1, v_1)} \quad (62)$$

The proofs of the relations (60), (61), and (62) are elementary and can be found in Björstad and Mandel (1991), Axelsson and Vassilevskii (1989), and Haase, Langer and Meyer (1991) respectively. Relation (62) means that γ coincides with the maximal eigenvalue of the generalized eigenvalue problem

$$(V_1^T K V_2)(V_2^T K V_2)^{-1}(V_2^T K V_1)v_1 = \lambda (V_1^T K V_1)v_1 \quad (63)$$

The error iteration schemes $z^{n+1} = E z^n$ corresponding to the exact ASM ($\tau = 1$) and the exact MSM are illustrated at the left and right sides of Figure 2, respectively. This figure shows that the exact MSM converges twice as fast as the corresponding ASM. More precisely, the following theorem holds.

Theorem 1. *Assume the splitting (57) with $\gamma \in [0, 1]$ defined by (58). Then the exact MSM converges in the energy norm with the rate γ^2 , that is,*

$$\|u - u^{n+1}\|_a \leq \gamma^2 \|u - u^n\|_a, \quad n = 1, 2, \dots \quad (64)$$

provided that $u^1 \in V$ is chosen such that $z^1 = (I - P_2)z^0 \in V_2^\perp$ (initial orthoprojection step onto V_2). The convergence rate of the corresponding exact ASM with $\tau = \tau_{\text{opt}} = 1$ is only γ , that is,

$$\|u - u^{n+1}\|_a \leq \gamma \|u - u^n\|_a, \quad n = 0, 1, 2, \dots \quad (65)$$

In this case, the ASM preconditioner C has the form

$$C = V^{-T} \begin{pmatrix} V_1^T K V_1 & 0 \\ 0 & V_2^T K V_2 \end{pmatrix} V^{-1} \quad (66)$$

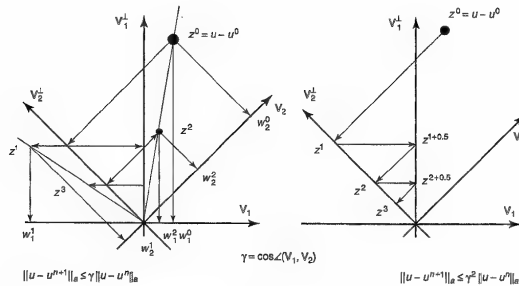


Figure 2. ASM and MSM corresponding to the splitting (57).

with the regular $N \times N$ basis transformation matrix $V = (V_1, V_2)$, and satisfies the spectral equivalence inequalities (39) with the sharp spectral equivalence constants $\underline{\gamma} = 1 - \gamma$ and $\bar{\gamma} = 1 + \gamma$.

The following straightforward spectral equivalence estimate for the inexact ASM preconditioner can be very useful in practice.

Corollary 1. *Let us assume that there are SPD subspace preconditioners C_1 and C_2 such that the spectral equivalence inequalities*

$$\underline{\gamma}_j C_j \leq V_j^T K V_j \leq \bar{\gamma}_j C_j \quad (67)$$

hold with positive $\underline{\gamma}_j$ and $\bar{\gamma}_j$ for $j = 1, 2$. Then the inexact ASM preconditioner

$$C = V^{-T} \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix} V^{-1} \quad (68)$$

is spectrally equivalent to K in the sense of the spectral equivalence inequalities (39) with the spectral equivalence constants $\underline{\gamma} = \min\{\underline{\gamma}_1, \underline{\gamma}_2\}(1 - \gamma)$ and $\bar{\gamma} = \max\{\bar{\gamma}_1, \bar{\gamma}_2\}(1 + \gamma)$.

Corollary 1 implies that the inexact ASM converges for $\tau \in (0, 2/\bar{\gamma})$ with the rate given by (38). The inexact sMSM version is discussed in Haase and Langer (1992). Further results for the inexact ASM, MSM, and sMSM versions follow from the general case discussed in the next section. We additionally refer to Aronszjan (1950) and Björstad and Mandel (1991) for the special case of the splitting into two subspaces, including the case where $V_1 \cap V_2$ is nontrivial like in the classical alternating Schwarz method introduced in Section 3.

3.3.2 The general case: splitting into J subspaces

Let us first consider the inexact ASM and the inexact ASM preconditioner introduced in Section 3.2.1. It was pointed out in this section that the convergence analysis (cf. rate estimate (38)) aims at the spectral estimates (39)–(41) of the ASM preconditioned matrix $C^{-1}K$, or equivalently, of the ASM operator \tilde{P} . Defining the so-called splitting norm

$$\|v\|^2 = \inf_{v = \sum_{j=1}^J v_j} \sum_{j=1}^J a_j(v_j, v_j) \quad (69)$$

where the infimum is taken over all possible splittings $v = \sum_{j=1}^J v_j$ of v with $v_j \in V_j$, we obtain the following exact representation of the minimal and maximal eigenvalues of $C^{-1}K$ resp. \tilde{P} .

Theorem 2.

$$\lambda_{\min}(C^{-1}K) = \lambda_{\min}(\tilde{P}) = \min_{v \in V} \frac{a(v, v)}{\|v\|^2}$$

and

$$\lambda_{\max}(C^{-1}K) = \lambda_{\max}(\tilde{P}) = \max_{v \in V} \frac{a(v, v)}{\|v\|^2} \quad (70)$$

Proof. follows from (39)–(41) and the observation that $\|v\| = a(\tilde{P}^{-1}v, v)$ (see e.g. Björstad and Mandel, 1991; Xu, 1992; and Oswald, 1994). Theorem 2 is closely related to the so-called fictitious space lemma that has been used in the Russian literature since the eighties (Matsokin and Nepomnyaschikh, 1985; Nepomnyaschikh, 1990) (see also Oswald (1994) and Griebel and Oswald (1995) for this relation). \square

The following two corollaries immediately follow from Theorem 2 and provide powerful tools for estimating (calculating) $\lambda_{\min}(\tilde{P})$ and $\lambda_{\max}(\tilde{P})$.

Corollary 2 (Lions' lemma, 1988). *Assume that there exists a positive constant c_L such that for all $v \in V$ there exists at least one splitting $v = \sum_{j=1}^J v_j$ such that the inequality*

$$\sum_{j=1}^J a_j(v_j, v_j) \leq c_L^2 a(v, v) \quad (71)$$

holds. Then

$$\lambda_{\min}(C^{-1}K) = \lambda_{\min}(\tilde{P}) \geq 1/c_L^2 \quad (72)$$

Proof. See also the original paper by Lions (1988). \square

Corollary 3 (subspace interaction lemma). *Let us define the $J \times J$ subspace interaction matrix $\Gamma = (\gamma_{ij})_{i,j=1,\dots,J}$ with the coefficients γ_{ij} stemming from the supremum (cf. also (62))*

$$\gamma_{ij} = \sup_{v_i \in V_i, v_j \in V_j} \frac{a(v_i, v_j)}{\sqrt{a_i(v_i, v_i)} \sqrt{a_j(v_j, v_j)}} \quad (73)$$

or the generalized strengthened Cauchy inequalities

$$|a(v_i, v_j)| \leq \gamma_{ij} \sqrt{a_i(v_i, v_i)} \sqrt{a_j(v_j, v_j)} \quad \forall v_i \in V_i, \forall v_j \in V_j \quad (74)$$

Then

$$\lambda_{\max}(C^{-1}K) = \lambda_{\max}(\tilde{P}) \leq \rho(\Gamma) \quad (75)$$

where $\rho(\Gamma)$ denotes the spectral radius of the subspace interaction matrix Γ .

Proof. See Dryja and Widlund (1995). \square

The convergence analysis of inexact multiplicative or hybrid versions of the Schwarz methods is more complicated. First of all, we need the so-called subspace contraction condition stating that the inequalities

$$a_j(\tilde{P}_j v_j, v_j) \leq \omega a_j(v_j, v_j) \quad \forall v_j \in V_j, \quad \forall j = 1, 2, \dots, J \quad (76)$$

hold for some constant $\omega \in (0, 2)$. Indeed, condition (76) ensures the contraction of the operator $I - \tilde{P}_j$ on the subspace V_j . We present here only two results concerning the inexact MSM and the inexact sMSM introduced in Section 3.2.2. For more results, we refer the reader to special papers on this topic, for example, Xu (1992), Dryja and Widlund (1995), and Griebel and Oswald (1995).

The sMSM produces a SPD preconditioner $C = K(I - E)^{-1}$ that can be used in the PCG method for solving our FE system (23). The following theorem again provides bounds for the minimal and maximal eigenvalues of the preconditioned matrix $C^{-1}K$.

Theorem 3. *Let us assume that the space splitting is stable in the sense of Corollaries 2–3 and that the subspace contraction condition (76) holds for some constant $\omega \in (1, 2)$. Then the spectral estimates*

$$\lambda_{\min}(C^{-1}K) = \lambda_{\min}(I - E) \geq \frac{2 - \omega}{\omega \rho^2(\Gamma) c_L^2}$$

and

$$\lambda_{\max}(C^{-1}K) = \lambda_{\max}(I - E) \leq 1 \quad (77)$$

hold with c_L and $\rho(\Gamma)$ defined in Corollaries 2 and 3 respectively.

Proof. See Smith, Bjørstad and Gropp (1996) for a slightly more general case, or the original paper by Dryja and Widlund (1995). \square

The following theorem gives an exact representation of the convergence rate $\|E\|_a$ of the MSM in the energy norm, where E again denotes the MSM error propagation (iteration) operator defined by (46):

Theorem 4. *Let us again assume that the subspace contraction condition (76) holds for some constant $\omega \in (0, 2)$. Then the energy norm of the MSM error propagation operator E resp. of the MSM iteration matrix E can be represented in the form*

$$\|E\|_a = \|E\|_K = q := \frac{c}{1+c} < 1 \quad (78)$$

where

$$c = \sup_{\|v\|_a=1} \inf_{v \in \sum_{j=1}^J V_j} \sum_{i=1}^J a(\tilde{P}_i \tilde{P}_i^{-1} \tilde{P}_i v_i, v_i) < \infty \quad (79)$$

with $w_i = \sum_{j=1}^J v_j - \tilde{P}_i^{-1} v_i$ and $\tilde{P}_i = \tilde{P}_i^* + \tilde{P}_i - \tilde{P}_i^* \tilde{P}_i = 2\tilde{P}_i - \tilde{P}_i^2$ ($\tilde{P}_i = \tilde{P}_i^*$).

Proof. See Xu and Zikatanov (2002) for the more general case of infinite-dimensional Hilbert spaces and of nonsymmetric, but elliptic bilinear forms. \square

In the case of the exact MSM, the representation can be simplified because $\tilde{P}_i = P_i = P_i^* = P_i^2$ is an orthoprojection with respect to the energy inner product $a(\cdot, \cdot)$. More precisely, the constant c in Theorem 4 can directly be rewritten in the form

$$c = \sup_{\|v\|_a=1} \inf_{v \in \sum_{j=1}^J V_j} \sum_{i=1}^J \sum_{j=1}^J \sum_{k=1}^J v_j \|v_k\|_a < \infty \quad (80)$$

Moreover, Theorem 4 immediately yields the convergence rate estimate

$$\begin{aligned} \|E_{\text{MSM}}\|_a &= \|E_{\text{MSM}}^* E_{\text{MSM}}\|_a \leq \|E_{\text{MSM}}\|_a \|E_{\text{MSM}}\|_a \\ &= \|E_{\text{MSM}}\|_a^2 = q^2 < 1 \end{aligned} \quad (81)$$

for sMSM, where E_{MSM} and E_{sMSM} denote the error propagation operators corresponding to the MSM and sMSM, respectively. Estimate (81) implies the spectral equivalence inequalities

$$(1 - q^2) C \leq K \leq 1 C \quad (82)$$

for the sMSM preconditioner $C = K(I - E_{\text{sMSM}})^{-1}$. The abstract representations and estimates given above are very essential for obtaining spectral equivalence or convergence rate estimates for specific subspace correction methods in concrete applications.

4 OVERLAPPING DOMAIN DECOMPOSITION METHODS

4.1 Basic construction principles and algorithms with generous overlap

As explained in Sections 2 and 3, the overlapping DD methods have a long history and can be completely treated within the framework of the Schwarz theory. More precisely,

- the splitting of (FE) space $V = \sum_{j=1}^J V_j$,
- the subspace bilinear forms $a_j(\cdot, \cdot)$, and

- the arrangement of the projection-like operations (additive, multiplicative, hybrid)

completely define the Schwarz method (preconditioner), the complete analysis of which is also covered by the Schwarz theory presented in Section 3.3.

Without loss of generality, we restrict ourselves to the (exact) additive version (preconditioner) that is the most important one in parallel computing. Then the analysis can mainly be reduced to the verification of the conditions formulated in Corollaries 2 and 3, namely,

- the verification of the stability (71) of the space splitting and
- the calculation of the subspace interaction measure $\rho(\Gamma)$.

For definiteness, we consider the heat conduction problem described by Example 1 as a model problem and assume a moderate and smooth behavior of the heat conduction coefficients. In the overlapping DD method, the splitting $V = \sum_{j=1}^J V_j$ of a (FE) space $V = V_h(\Omega) \subset H_0^1(\Omega)$ corresponds to an overlapping DD $\Omega = \bigcup_{j=1}^J \Omega_j$ of the computational domain Ω , where the subspaces usually have the form $V_j = V \cap H_0^1(\Omega_j)$, that is, the subspace problems are local Dirichlet FE problems. There are several methods for constructing overlapping domain decompositions. Let us mention at first two simple techniques that both start from a coarse shape-regular conform triangulation T_H of $\bar{\Omega} = \bigcup_{j=1}^J \bar{\Omega}_{H,j}$ and proceed with its refinement, resulting in a shape-regular conform fine grid discretization T_h of $\bar{\Omega} = \bigcup_{j=1}^J \bigcup_{i \in \mathcal{T}_{h,j}} \bar{\Omega}_{h,i,j} = \bigcup_{i \in \mathcal{T}_h} \bar{\Omega}_{h,i}$. For definiteness, we assume that these triangulations are provided by triangles ($d = 2$) or tetrahedra ($d = 3$), and, for simplicity, we also assume that linear elements are used for generating the FE spaces. The parameters H and h stand for the typical sizes of the coarse and the fine quasisuniform triangulations respectively. Now we associate with each coarse grid vertex $x^{(j)}$ ($j = 1, 2, \dots, J$) some subdomain $\bar{\Omega}_j$ that is built by all coarse simplices containing $x^{(j)}$ as a vertex. This gives our first overlapping domain decomposition (ODD1) of Ω , where the overlap $\delta = O(H)$. Another one is given by associating with each coarse grid simplex $\bar{\tau}_{H,j}$ some subdomain $\bar{\Omega}_j$ that is built by this simplex and all simplices touching this simplex at least in one vertex, where $j = 1, 2, \dots, J$ with $J = \tilde{J}$. This again gives us an overlapping domain decomposition (ODD2) of Ω , where the overlap $\delta = O(H)$. Several generalizations of these techniques are feasible. For instance, one can first build a nonoverlapping domain decomposition with subdomains $\bar{\Omega}_j$ consisting of one or several coarse grid elements and then extend them by adding some layers of fine grid elements around these subdomains, giving the subdomains Ω_j of an

overlapping domain decomposition (ODD(3)), where δ is the thickness of the layers, that is, the overlap. Thus, ODD2 is a special ODD(H) method.

Let C be the (exact) additive Schwarz preconditioner (cf. (36))

$$C^{-1} = \sum_{j=1}^J V_j (V_j^T K V_j)^{-1} V_j^T = \sum_{j=1}^J V_j K_j^{-1} V_j^T \quad (83)$$

corresponding to the space splitting

$$V = V_h = \sum_{j=1}^J V_j, \quad V_j = V_h \cap H_0^1(\Omega_j) = \text{span} \Phi V_j \quad (84)$$

that is based on one of the overlapping domain decompositions $\bigcup_{j=1}^J \Omega_j$ of Ω with $O(H)$ overlap as described above, where the $N \times N_j$ matrix V_j picks exactly those basis functions from the fine grid nodal basis Φ , which belong to the inner nodes in Ω_j . The SPD $N_j \times N_j$ matrix K_j is nothing but the stiffness matrix belonging to the local FE Dirichlet problem in Ω_j . Using the general Schwarz theory, we can prove that

$$\kappa(C^{-1}K) = O(H^{-2}) \quad (85)$$

that is, owing to the generous $O(H)$ overlap, the relative spectral condition number does not depend on the fine grid discretization parameter h in a bad way, but on the domain decomposition parameter H . This is totally unacceptable for the use of this preconditioner in a massively parallel solver environment. The bad dependence on H is due to the absence of some coarse grid solver managing the global information transport that is essential for elliptic problems (see Widlund (1988) and Smith, Bjørstad and Gropp (1996) for a more detailed discussion of this issue in connection with DD methods).

There are a lot of possibilities to include such mechanisms for global information exchange into the Schwarz preconditioner. For the overlapping domain decompositions presented above, the natural way certainly consists in adding the coarse grid FE space $V_0 = V_h = \text{span} \Phi V_0 \subset V \subset H_0^1(\Omega)$ to the splitting (84), that is,

$$V = V_h = V_0 + \sum_{j=1}^J V_j = \sum_{j=0}^J V_j \quad (86)$$

Now, the corresponding two-level (coarse level and fine level) overlapping ASM preconditioner

$$C^{-1} = \sum_{j=0}^J V_j K_j^{-1} V_j^T \quad (87)$$

gives an optimal relative spectral condition number estimate.

Theorem 5. *The exact two-level ASM preconditioner (87) based on an overlapping domain decomposition with a uniform overlap width $O(H)$ provides an optimal preconditioner in the sense that $\kappa(C^{-1}K) < c$, where the positive constant c does not depend on h , H , and J .*

Proof. was given by Dryja and Widlund (1989) on the basis of the Schwarz theory and some technical lemmas (partition of unity, stability of L_2 -projection in H^1) (see also Smith, Björstad and Gropp, 1996). \square

The theorem remains obviously true for the inexact version where the local stiffness matrices K_j are replaced by suitable spectrally equivalent preconditioners C_j for $j = 1, 2, \dots, J$. The analysis of multiplicative versions follows the same line of the general Schwarz theory (see e.g. Bramble *et al.*, 1991; Xu, 1992; and Xu and Zikatanov, 2002). The results can be extended to coarse grid spaces V_H , which are not subspaces of the FE fine grid space $V = V_h$.

However, there are two drawbacks of two-level ASM preconditioners with a generous overlap. The first problem is connected with jumps in the coefficients. In contrast to the nonoverlapping DD methods (cf. Section 5), the influence of the jumps in the coefficients with large jumps is still not completely understood. The second problem is connected with the influence of the overlap width δ on $\kappa(C^{-1}K)$. It is clear that the larger the overlap, the more expensive are the local problems that we have to solve. The computational overhead becomes significant when h becomes small with respect to H . We discuss this problem in the next section.

4.2 Domain decomposition algorithms with a small overlap

In the case of a small overlap δ , Dryja and Widlund (1994) proved the following theorem.

Theorem 6. *The exact two-level ASM preconditioner (87) based on an overlapping domain decomposition with a uniform overlap width $O(\delta)$ gives the estimate*

$$\kappa(C^{-1}K) \leq \bar{c} \left(1 + \frac{H}{\delta}\right) \quad (88)$$

of the relative spectral condition number $\kappa(C^{-1}K)$, where the positive constant \bar{c} does not depend on h , H , J , and δ .

Brenner (2000) showed that this result is sharp in the case of minimal overlap ($\delta = h$), that is, there exists a positive

constant \bar{c} that is independent of h , H , and J such that $\kappa(C^{-1}K) \geq \bar{c}(H/h)$. In the same paper, she proved that $\kappa(C^{-1}K) = O((H/h)^2)$ in the case of fourth-order elliptic boundary value problems.

Therefore, a small overlap really affects the preconditioning effect of the two-level ASM preconditioner (87) in a very bad way. On the other hand, the $O(H)$ overlap means that additional $O((H/h)^d)$ unknowns are added to the local problems in contrast to $O((H/h)^{d-1})$ unknowns in the case of an $O(h)$ overlap. Bank *et al.* (2002) have recently proposed a two-level hierarchical overlapping ASM preconditioner that adds only $O((H/h)^{d-1})$ unknowns to the local problems as in the case of small overlap, and that results in a uniformly bounded relative condition number estimate, as in the case of Theorem 6.

4.3 Multilevel versions

The two-level Schwarz methods, described above, use a fine (h) and a coarse (H) mesh capturing the local (high-frequency) and the global (low-frequency) parts in the solution (iteration error) respectively. This approach is satisfactory if efficient local and global solvers (preconditioners) are available. However, if the global problem is large, that is, H is relatively small, then we can again apply a two-level algorithm to the coarse grid problem using again some coarser grid. The recursive application to the coarse grid problems results in a multilevel ASM preconditioner. To be more precise, we assume that the coarse grid (triangulation) $T_0 = T_H$ is refined L times giving the finer and finer triangulations T_1, \dots, T_{L-1} , and $T_L = T_h$. For each level $l = 1, 2, \dots, L$ of the triangulations, with the exception of the coarsest level $l = 0$, we construct some overlapping domain decomposition $\Omega = \bigcup_{j=1}^J \Omega_{l,j}$ and connect with this multilevel overlapping domain decomposition the multilevel splitting of the FE space

$$V = V_h = V_0 + \sum_{l=1}^L \sum_{j=1}^J V_{l,j} \quad (89)$$

in the same way as above, where the subspaces $V_{l,j} = \text{span} \Phi_{l,j}$ can again be generated by using the $N \times N_{l,j}$ basis transformation matrices $V_{l,j}$. Now the corresponding (exact) multilevel overlapping ASM preconditioner can be written in the form

$$C^{-1} = V_0 K_0^{-1} V_0^T + \sum_{l=1}^L \sum_{j=1}^J V_{l,j} K_{l,j}^{-1} V_{l,j}^T \quad (90)$$

Let us consider one extreme case where the subdomains $\Omega_{l,j}$ are simply the supports of the nodal basis functions

$\phi_{l,j}$ belonging to the node $x_{l,j}$ in the l -level triangulation T_l , that is, now the subspaces $V_{l,j} = \text{span}(\phi_{l,j}) = \text{span} \Phi_{l,j}$ are one-dimensional, where $\Phi = \Phi_h = \Phi_L = [\phi_{L,j}]$ denotes the fine grid basis and the $N \times 1$ matrix $V_{l,j}$ provides the representation of the basis function $\phi_{l,j}$ in the fine grid basis. We mention that this overlapping DD is closely related to ODD1, but now the elements around the node $x_{l,j}$ are taken from the triangulation T_l and not T_{l-1} . The exact multilevel overlapping ASM preconditioner corresponding to this overlapping domain decomposition is called *multilevel diagonal scaling* (MDS) preconditioner and was introduced by Zhang (1992). Since for second-order elliptic problems the 1×1 matrices $K_{l,j}$ obviously behave like h_l^{d-2} , the inexact version of the MDS preconditioner can be written in the form

$$C^{-1} = V_0 K_0^{-1} V_0^T + \sum_{l=1}^L h_l^{2-d} \sum_{j=1}^J V_{l,j} V_{l,j}^T \quad (91)$$

The inexact multilevel overlapping ASM preconditioner (91) was first proposed by Bramble, Pasciak and Xu (1990) and is nowadays known as BPX preconditioner.

Theorem 7. *The MDS and the BPX preconditioners are optimal preconditioners in the sense that there is some positive constant c that does not depend on h and L such that the relative spectral condition number $\kappa(C^{-1}K) \leq c$, and the arithmetical cost $\text{ops}(C^{-1}d)$ for the preconditioning operation is proportional to the number of unknowns $N_h = O(h^{-d})$ on the finest grid.*

Proof. The original proof by Bramble, Pasciak and Xu (1990) provided weaker (nonoptimal) bounds depending on the number of refinement levels $L = O(\log(1 + (H/h)))$ (see also Zhang (1992) for the MDS preconditioner). The optimality of the BPX preconditioner was first proved by Oswald (1992) using Besov space techniques (see also Oswald, 1994). \square

Multiplicative and hybrid versions of these multilevel Schwarz methods are closely related to multigrid methods, which are discussed in Chapter 20, this Volume; see also Bramble and Zhang (2000) for this relation.

5 NONOVERLAPPING DOMAIN DECOMPOSITION METHODS

5.1 Iterative substructuring methods

For definiteness, we consider the heat conduction problem described by Example 1 as model problem. In practice, the

heat conduction coefficient $\alpha(\cdot)$ typically has jumps due to different materials. As in Section 2, we assume that the computational domain

$$\bar{\Omega} = \bigcup_{j=1}^J \bar{\Omega}_j \quad (92)$$

is decomposed into J nonoverlapping subdomains in such a way that the coefficient jumps are along with the boundary of the subdomains. For simplicity, we assume that in each subdomain Ω_j the coefficient $\alpha(\cdot)$ has the constant positive value α_j . Further, we assume that the domain decomposition is quasiregular in the sense that the subdomains are images of some reference domain, or a few reference domains, by a quasiregular mapping with the scaling H . Therefore, H can be viewed as typical subdomain diameter such that $J = O(H^{-d})$. As described in Section 2, we provide every subdomain with a triangulation T_j such that the triangulation T_h of the total computational domain $\bar{\Omega} = \bigcup_{j=1}^J \bigcup_{\tau \in T_j} \tau$ is conform and quasiregular in the sense that there is a typical element diameter h such that $N = N_h = O(h^{-d})$. Thus, the number of the internal subdomain unknowns N_{hj} behaves like $O((H/h)^d)$. The FE discretization with the arrangement (12) of the FE basis Φ leads to the block structure (13) of the FE equations. The stiffness matrix K and the load vector f can obviously be represented in the form

$$K = \sum_{j=1}^J A_j^T K_j A_j \quad \text{and} \quad f = \sum_{j=1}^J A_j^T f_j \quad (93)$$

where the $N_j \times N$ Boolean subdomain connectivity matrices A_j are mapping some vector $u \in \mathbb{R}^N$ of all nodal values onto the vector $u_j = A_j u \in \mathbb{R}^{N_j}$ of the subdomain nodal values. The $N_j \times N_j$ subdomain stiffness matrices K_j and the subdomain load vectors $f_j = A_j f \in \mathbb{R}^{N_j}$ can be structured in the same way as we have structured K and f in (13), that is,

$$A_j = \begin{pmatrix} A_{C_j} & A_{C_j H_j} \\ A_{H_j C_j} & A_{H_j H_j} \end{pmatrix}, \quad K_j = \begin{pmatrix} K_{C_j C_j} & K_{C_j H_j} \\ K_{H_j C_j} & K_{H_j H_j} \end{pmatrix}, \quad f_j = \begin{pmatrix} f_{C_j} \\ f_{H_j} \end{pmatrix} \quad (94)$$

The matrices K_{H_j} correspond to the local homogeneous Dirichlet problems, whereas the matrices K_{C_j} arise from the FE discretization of the local Neumann problems, at least, for the subdomains with $\partial\Omega_j \cap \partial\Omega = \emptyset$. For our model problem, these matrices are singular, where the kernel (null space) $\ker(K_j) = \text{span}(1_j)$ is spanned by $1_j = (1, 1, \dots, 1) \in \mathbb{R}^{N_j}$.

The block Gaussian elimination of the internal subdomain unknowns \mathbf{u}_i reduces the solution of the FE equation (13) to the solution of the Schur-complement problem

$$\mathbf{S}_C \mathbf{u}_C = \mathbf{g}_C \quad (95)$$

that was explicitly formed and directly solved in the classical substructuring Algorithm 3. The Schur complement \mathbf{S}_C and the right-hand side \mathbf{g}_C can be assembled from the local Schur complements \mathbf{S}_{C_j} and the local right-hand sides \mathbf{g}_{C_j} in the same way as \mathbf{K} and \mathbf{f} were assembled in (93) from \mathbf{K}_j and \mathbf{g}_j respectively, that is,

$$\mathbf{S}_C = \sum_{j=1}^J \mathbf{A}_{C_j}^T \mathbf{S}_{C_j} \mathbf{A}_{C_j} \quad \text{and} \quad \mathbf{g}_C = \sum_{j=1}^J \mathbf{A}_{C_j}^T \mathbf{g}_{C_j} \quad (96)$$

As mentioned in Section 2, the iterative solution of (95) avoids the very expensive forming of the Schur complement \mathbf{S}_C . Solving (95) by the PCG methods lead to our first iterative substructuring method called Schur-complement CG. In each iteration step of the Schur complement CG, we need one matrix-by-vector multiplication of the form

$$\begin{aligned} \mathbf{S}_C \mathbf{v}_C^a &= \sum_{j=1}^J \mathbf{A}_{C_j}^T \mathbf{S}_{C_j} \mathbf{A}_{C_j} \mathbf{v}_C^a \\ &= \sum_{j=1}^J \mathbf{A}_{C_j}^T (\mathbf{K}_{C_j} - \mathbf{K}_{C_j} \mathbf{K}_j^{-1} \mathbf{K}_{j,C_j}) \mathbf{A}_{C_j} \mathbf{v}_C^a \end{aligned} \quad (97)$$

requiring the direct solution of J systems (local Dirichlet problems)

$$\mathbf{K}_{j,C_j} \mathbf{w}_C^a = \mathbf{K}_{j,C_j} \mathbf{A}_{C_j} \mathbf{v}_C^a, \quad j = 1, \dots, J \quad (98)$$

which can be done completely in parallel. Moreover, the factorization of the matrices \mathbf{K}_{j,C_j} in a preprocessing step and the use of sparse direct techniques can make this multiplication operation very efficient (see e.g. the classical monograph by George and Liu (1981) and more recent papers by Demmel *et al.* (1999) and Gupta (2002)). Nevertheless, for real large-scale problems, this operation is a bottleneck of the Schur-complement CG. The use of inexact (iterative) solvers for the local Dirichlet problems (98) is dangerous. We discuss the naive use of inexact solvers in Section 5.3.

The forming of the Schur complement is some kind of a preconditioning operation. Indeed, the spectral condition number of the Schur complement $\kappa(\mathbf{S}_C) = \lambda_{\max}(\mathbf{S}_C) / \lambda_{\min}(\mathbf{S}_C) = O(H^{-1}h^{-1})$ is much better than the spectral condition number of the original stiffness matrix $\kappa(\mathbf{K}) = O(h^{-2})$ since $h \ll H$ (see e.g. Brenner, 1999).

However, $\kappa(\mathbf{S}_C)$ still depends on the DD parameter H and the global discretization parameter h as well as on the coefficient jumps in a bad way. Thus, we need a SPD Schur-complement preconditioner \mathbf{C}_C removing the influence of all these parameters in such a way that $\kappa(\mathbf{C}_C^{-1} \mathbf{S}_C)$ does not depend too much on these parameters under the restriction that the preconditioning operation $\mathbf{w}_C^a = \mathbf{C}_C^{-1} \mathbf{d}_C^a$, mapping the defect \mathbf{d}_C^a into preconditioned defect \mathbf{w}_C^a , is sufficiently cheap. Since the Schur-complement preconditioner is one of the most important ingredients of many iterative substructuring methods, we discuss this topic in more detail in the next section.

In this chapter, we only consider iterative DD methods for h -versions of the FEM. The hp methods, by which we here mean both finite element and spectral element methods, are gaining growing attention due to the ability to attain exponential convergence even for problems with singularities. Although this noticeable advantage is often damaged by the high cost of the setup procedure caused by the high fill-in of the stiffness matrices and complex algorithms for calculating their entries, the computational cost grows with p algebraically at the worst. Therefore, there is a strong incentive to achieve a better performance by means of smart hp algorithms. The literature on hp methods is very vast. We refer the reader to Chapter 5, Chapter 6 of this Volume and Chapter 3, Volume 3 for more information. In spite of the high interest, the toolkit of fast solvers for the systems arising from hp discretizations is much smaller than that for the h -version. In the last decade, the major progress in this area has been achieved on the basis of DD approaches. The formation of the basic features of hp DD algorithms is due to the contributions by Babuska, Craig, Mandel and Pitkäranta (1991), Pavarino (1994), Ivanov and Korneev (1995), Ainsworth (1996), Ivanov and Korneev (1996), Pavarino and Widlund (1996), Widlund (1996), Ainsworth and Senior (1997), Casarin (1997), Oden, Patra and Feng (1997), and Korneev and Jensen (1997). Let us mention that the spectrally equivalent finite-difference-like preconditioners for the local problems were suggested and studied by Orzag (1980), Bernardi, Dauge and Maday (1992), and Casarin (1997) for spectral discretizations, and by Ivanov and Korneev (1996) and Korneev and Jensen (1999) for hierarchical discretizations. Later studies paid more attention to a more elaborate design of all components of DD algorithms, that is, fast solvers for the local Dirichlet problems, efficient prolongations from the edges in 2D and from the faces in 3D cases, respectively, edge and face Schur-complement preconditioners, solvers for the wire-basket problem, and so on. In this relation, we refer to Korneev (2001, 2002a,b), Beuchler (2002), Korneev *et al.* (2002), and Korneev, Langer and Xanthis (2003). These studies resulted in fast

Dirichlet DD preconditioners for second-order elliptic equations. Some useful properties have been added to hp DD methods by the use of nonconforming discretizations and, in particular, by the mortar and FETI methods (see e.g. Bernardi, Maday and Sacchi-Landriani, 1989; Bernardi and Maday and Patera, 1993; Ben Belgacem and Maday, 1999; Bernardi, Maday and Sacchi-Landriani, 1989). Note that the components of the fast solvers developed for the Dirichlet DD preconditioners may be applied to these discretizations as well.

5.2 Schur-complement preconditioners

In this section, we look for SPD Schur-complement preconditioners \mathbf{C}_C satisfying the following conditions:

1. Spectral equivalence condition: the spectral equivalence inequalities

$$\underline{\gamma}_C \mathbf{C}_C \leq \mathbf{S}_C \leq \bar{\gamma}_C \mathbf{C}_C \quad (99)$$

should hold with positive spectral equivalence constants $\underline{\gamma}_C$ and $\bar{\gamma}_C$ such that $\kappa(\mathbf{C}_C^{-1} \mathbf{S}_C) \leq \bar{\gamma}_C / \underline{\gamma}_C$ does not, or only weakly depend on h , H , and the coefficient jumps. The latter property is sometimes also called robustness with respect to coefficient jumps.

2. Efficiency condition: the number of arithmetical operations $\text{ops}(\mathbf{C}_C^{-1} \mathbf{d}_C^a)$ needed for the preconditioning operation should be of the order $O(N_C) \dots O(N)$, or at least should not disturb the overall complexity of the algorithm too much.

3. Parallelizability condition: the preconditioning operation $\mathbf{C}_C^{-1} \mathbf{d}_C^a$ should not disturb the numerical and parallel efficiency of the total algorithm too much. However, we should be aware that in many Schur-complement preconditioners some coarse grid solver managing the global information transport is hidden. Thus, the coarse grid solver requires global communication and is some bottleneck in the parallelization.

In the literature, there are several basic proposals for Schur-complement preconditioners. Many of them are based on the fact that the Schur-complement energy $(\mathbf{S}_C \mathbf{u}_C, \mathbf{u}_C)$ is spectrally equivalent to the broken-weighted $H^{1/2}$ -norm

$$\begin{aligned} \|\mathbf{u}_C\|_{H^{1/2}(\Gamma_C)}^2 &:= \sum_{j=1}^J \alpha_j |\mathbf{u}_{C_j}|_{H^{1/2}(\Gamma_j)}^2 \\ &= \sum_{j=1}^J \alpha_j \int_{\Gamma_j} \int_{\Gamma_j} \frac{|\mathbf{u}_{C_j}(\mathbf{y}) - \mathbf{u}_{C_j}(\mathbf{x})|^2}{|\mathbf{y} - \mathbf{x}|^d} d\mathbf{s}_y d\mathbf{s}_x \end{aligned} \quad (100)$$

with $\Gamma_i = \partial\Omega_i$, that is, there exist positive constants $\underline{\delta}_C$ and $\bar{\delta}_C$, which are independent of h , H , and the coefficient jumps, such that

$$\begin{aligned} \underline{\delta}_C \|\mathbf{u}_C\|_{H^{1/2}(\Gamma_C)}^2 &\leq (\mathbf{S}_C \mathbf{u}_C, \mathbf{u}_C) \\ &\leq \bar{\delta}_C \|\mathbf{u}_C\|_{H^{1/2}(\Gamma_C)}^2 \quad \forall \mathbf{u}_C \in \mathbb{R}^{N_C} \end{aligned} \quad (101)$$

Evidently,

$$\begin{aligned} \|\mathbf{u}_C\|_{\mathbf{S}_C}^2 &= (\mathbf{S}_C \mathbf{u}_C, \mathbf{u}_C) \\ &= \sum_{j=1}^J \alpha_j (\bar{\mathbf{S}}_{C_j} \mathbf{u}_{C_j}, \mathbf{u}_{C_j}) = \sum_{j=1}^J \alpha_j \inf_{\mathbf{v}_j|_{\Gamma_j} = \mathbf{u}_{C_j}} \|\nabla \mathbf{v}_j\|_{L_2(\Omega_j)}^2 \end{aligned} \quad (102)$$

where the subdomain Schur complements $\bar{\mathbf{S}}_{C_j}$ arise from the case $\alpha_j = 1$ and the infimum is taken over all FE functions $\mathbf{v}_j \in \mathbf{V}_j = \mathbf{V}|_{\bar{\Omega}_j}$ living in $\bar{\Omega}_j$ and coinciding with \mathbf{u}_{C_j} on Γ_j . Therefore, the equivalence (101) is the same as the equivalence of the infimum in (102) to the $H^{1/2}(\Gamma_j)$ -norm in (100) and requires the trace and lifting (prolongation) theorems for functions from the FE space. In simple cases, the left inequality in (101) is an obvious consequence of the trace theorem for functions from $H^1(\Omega_j)$. The right inequality in (101) requires some special proof, which, for example, may be found in Nepomnyashchikh (1991b).

Let us first describe Dryja's classical Schur-complement preconditioner that is just applicable to the L-shaped domain sketched in Figure 1. Here, the interface Γ_C consists only of one straight piece with N_C (inner) nodal points. For this simple but characteristic example, Dryja (1982) proposed the preconditioner

$$\begin{aligned} \mathbf{C}_C &= \mathbf{B}_C^{1/2} \\ &= \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}^{1/2} = \mathbf{F}_C^T \mathbf{A}_C^{1/2} \mathbf{F}_C \end{aligned} \quad (103)$$

that is nothing but the square root of the scaled discretized 1D Laplacian \mathbf{B}_C along the interface under homogeneous Dirichlet boundary conditions at the end points of the interface. Dryja (1982) proved the spectral equivalence inequalities (99) with h -independent spectral equivalence constants $\underline{\gamma}_C$ and $\bar{\gamma}_C$ (see also theorem 1 in Dryja, 1984). It is well known that \mathbf{B}_C has the eigenvalues $\lambda_k = 4 \sin^2(k\pi/(2(N_C+1)))$ and the corresponding orthonormal eigenvectors $\mathbf{v}_{C,k} = [\sqrt{2/(N_C+1)} \sin(k\pi l/(N_C+1))]_{l=1, \dots, N_C}$, where k is running from 1 to N_C . Therefore, in (103), $\mathbf{A}_C^{1/2} =$

$\text{diag}(\lambda_k^{1/2})$ and $\mathbf{F}_C = [\mathbf{v}_{C,1}, \dots, \mathbf{v}_{C,N_C}]$. Since the Fourier matrix \mathbf{F}_C is orthogonal, that is, $\mathbf{F}_C^{-1} = \mathbf{F}_C^T$, the preconditioning equation $\mathbf{C}_C \mathbf{w}_C = \mathbf{d}_C$ can be solved in the following three steps:

$$\begin{aligned} \text{Fourier analysis: } & \mathbf{x}_C = \mathbf{F}_C \mathbf{d}_C \\ \text{Diagonal scaling: } & \mathbf{y}_C = \mathbf{A}_C^{-1/2} \mathbf{x}_C \\ \text{Fourier synthesis: } & \mathbf{w}_C = \mathbf{F}_C^T \mathbf{y}_C \end{aligned}$$

The complexity of this preconditioning operation is dominated by the Fourier analysis and Fourier synthesis. Using the fast Fourier transformation (FFT), the total complexity of the preconditioning operation is of the order $N_C \ln(N_C)$. Dryja's preconditioner was improved by Golub and Mayers (1984), Björstam and Widlund (1986), Chan (1987), and others.

The construction of good Schur-complement preconditioners in the general case of many nonoverlapping subdomains is more involved. In the two-dimensional case, Bramble, Pasciak and Schatz (1986) proposed an ASM preconditioner of the form

$$\mathbf{C}_C^{-1} = \sum_E \mathbf{V}_E \mathbf{C}_E^{-1} \mathbf{V}_E^T + \sum_V \mathbf{V}_V \mathbf{C}_V^{-1} \mathbf{V}_V^T \quad (104)$$

in which the vertices are separated from the edges of the subdomains arising from some coarse grid domain decomposition. Here, \mathbf{V}_E takes the edge nodal basis functions (unknowns) belonging to the edge E from the nodal basis functions (unknowns), \mathbf{V}_V^T transforms the nodal basis functions into the coarse grid (hierarchical) vertex basis functions, \mathbf{C}_E is an edge Schur-complement preconditioner as discussed above (Dryja's type), and \mathbf{C}_V denotes a coarse grid preconditioner, for example, \mathbf{C}_V can be the coarse grid stiffness matrix. The latter one manages the global information exchange. Bramble, Pasciak and Schatz (1986) proved that the relative condition number $\kappa(\mathbf{C}_C^{-1} \mathbf{S}_C)$ grows at most like $(1 + \ln(H/h))^2$ as h tends to zero. Moreover, the right averaging of the coefficients of adjacent subdomains makes the BPS preconditioner, how the Schur-complement preconditioner (104) is nowadays called, quite robust with respect to coefficient jumps. In a series of papers, Bramble, Pasciak and Schatz (1986, 1987, 1988, 1989) generalized these ideas in many directions, including Schur-complement preconditioners for the 3D case.

Another type of Schur-complement preconditioners relies on the transformation of the nodal basis to a multilevel basis (generating system). This was very successful in constructing preconditioners for the finite element stiffness matrix \mathbf{K} (the hierarchical preconditioner of Yserantant and the BPS preconditioner). It is easy to see that the same transformations restricted to the coupling boundary (interface) nodes

result in Schur-complement preconditioners possessing at least the same quality as the corresponding preconditioners for \mathbf{K} . Smith and Widlund (1990) and Haase, Langer and Meyer (1991) proposed hierarchical Schur-complement preconditioners, which asymptotically behave like the BPS preconditioner in the 2D case. The BPX Schur-complement preconditioner was introduced by Tong, Chan and Kuo (1991). It is asymptotically optimal, but is sensitive to coefficient jumps.

There are a lot of other proposals for constructing Schur-complement preconditioners. Let us here mention only the wire-basket-based Schur-complement preconditioners introduced by Dryja, Smith and Widlund (1994), the probing technique proposed by Chan and Mathew (1994) (see also Keyes and Gropp, 1987), and the techniques borrowed from the boundary element method (see e.g. Carstensen, Kuhn and Langer, 1998; Haase *et al.*, 1997; and Steinbach, 2003).

Finally, we refer to the Neumann-Dirichlet and to the Neumann-Neumann preconditioners, which are special Schur-complement preconditioners approved to be very robust in practical applications. The Neumann-Neumann preconditioners are discussed in Section 5.4 in detail.

5.3 Inexact subdomain solvers

5.3.1 Effects of inexact subdomain solvers

Let us consider the discrete harmonic ($a(\cdot, \cdot)$ -harmonic) splitting

$$\mathbf{V} = \mathbf{V}_h = \mathbf{V}_C^* \oplus \mathbf{V}_I \quad (105)$$

of the FE space \mathbf{V} into the discrete harmonic space

$$\begin{aligned} \mathbf{V}_C^* &= P^* \mathbf{V}|_{\Gamma_C} = \{u \in \mathbf{V} : a(u, v) = 0 \forall v \in \mathbf{V}_I\} \\ &= \text{span} \Psi_C^* = \text{span} \Phi \mathbf{V}_C^* \end{aligned} \quad (106)$$

and the interior subdomain (bubble) space $\mathbf{V}_I = \text{span} \Phi \mathbf{V}_I$ and $\mathbf{V}_I \oplus \dots \oplus \mathbf{V}_{I_i}$, with the basis transformation matrices

$$\begin{aligned} \mathbf{V}_C^* &= \begin{pmatrix} \mathbf{I}_C \\ -\mathbf{K}_I^{-1} \mathbf{K}_{IC} \end{pmatrix}_{N \times N_C} \\ &= \begin{pmatrix} \mathbf{I}_C \\ \mathbf{P}_{IC}^* \end{pmatrix}_{N \times N_C} \quad \text{and} \quad \mathbf{V}_I = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_I \end{pmatrix}_{N \times N_I} \end{aligned} \quad (107)$$

where $\Psi_C^* = \Phi \mathbf{V}_C^*$ is called the discrete harmonic basis and P^* is the discrete harmonic extension (prolongation) operator mapping a FE function u_C living on Γ_C to some FE function $P^* u_C$ that coincides with u_C on Γ_C and is discrete harmonic in all subdomains Ω_i . Owing to the

construction of \mathbf{V}_C^* , the splitting is orthogonal with respect to the energy inner product $a(\cdot, \cdot)$. Therefore, the exact ASM preconditioner

$$\mathbf{C} = \begin{pmatrix} \mathbf{I}_C & \mathbf{K}_{CI} \mathbf{K}_I^{-1} \\ \mathbf{0} & \mathbf{I}_I \end{pmatrix} \begin{pmatrix} \mathbf{S}_C & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_I \end{pmatrix} \begin{pmatrix} \mathbf{I}_C & \mathbf{0} \\ \mathbf{K}_I^{-1} \mathbf{K}_{IC} & \mathbf{I}_I \end{pmatrix} = \mathbf{K} \quad (108)$$

must coincide with \mathbf{K} . This factorization of \mathbf{K} is used in many applications. Replacing \mathbf{S}_C by a Schur complement preconditioner \mathbf{C}_C (cf. Section 5.3), we arrive at a partially inexact ASM preconditioner that corresponds to the Schur complement iteration methods. Replacing additionally \mathbf{K}_I by some preconditioner \mathbf{C}_I gives us the full inexact ASM preconditioner. Owing to Corollary 1, both inexact ASM preconditioners are covered by the general theory with $\gamma = \cos \angle(\tilde{\mathbf{V}}_C, \mathbf{V}_I) = 0$. The naive replacement of \mathbf{K}_I^{-1} in the left and right factors of the factorization (108) is dangerous because it changes the angle between the subspaces. For instance, the 'inversion' of \mathbf{K}_I by one multigrid cycle, that is, replacement of \mathbf{K}_I^{-1} by $(\mathbf{I}_I - \mathbf{E}_I^T) \mathbf{K}_I^{-1} (\mathbf{I}_I - \mathbf{E}_I)$ ($s=1$), will, in general, not result in a stable splitting even if the multigrid convergence rate, that is, the energy norm of the multigrid iteration operator $|\mathbf{E}_I|_{\mathbf{K}_I}$, is less than 1 independently of the local discretization parameter H/h (as usual). One needs at least $s = O(\ln(H/h))$ multigrid cycles to get a stable extension (see e.g. Haase, Langer and Meyer, 1991). The next section shows that a stable splitting requires a careful choice of the extension operator \tilde{P} resp. $\tilde{\mathbf{P}}_{IC}$ replacing the discrete harmonic extension operator P^* resp. \mathbf{P}_{IC}^* .

5.3.2 The bounded extension splitting

Let us now consider the bounded extension splitting

$$\mathbf{V} = \mathbf{V}_h = \tilde{\mathbf{V}}_C \oplus \mathbf{V}_I \quad (109)$$

of the FE space \mathbf{V} in the direct sum of the former subdomain space \mathbf{V}_I that remains unchanged and the bounded extension space

$$\begin{aligned} \tilde{\mathbf{V}}_C &= \tilde{P} \mathbf{V}|_{\Gamma_C} = \{\tilde{P} u \in \mathbf{V} : u_C = u|_{\Gamma_C} \text{ given}\} \\ &= \text{span} \tilde{\Psi}_C = \text{span} \Phi \tilde{\mathbf{V}}_C \end{aligned} \quad (110)$$

with the basis transformation matrices

$$\tilde{\mathbf{V}}_C = \begin{pmatrix} \mathbf{I}_C \\ \tilde{\mathbf{P}}_{IC} \end{pmatrix}_{N \times N_C} : \mathbb{R}_{N_C} \rightarrow \mathbb{R}_N \quad (111)$$

We assume that there exists some constant $c_E^2 \geq 1$ independent of our bad parameters such that

$$\left\| \begin{pmatrix} \mathbf{I}_C \\ \tilde{\mathbf{P}}_{IC} \end{pmatrix} u_C \right\|_{\mathbf{K}} \leq c_E \|u_C\|_{\mathbf{S}_C} \quad \forall u_C \in \mathbb{R}^{N_C} \quad (112)$$

Together with the minimal energy property (102), inequality (108) gives the following spectral equivalence relations

$$\begin{aligned} (\mathbf{S}_C u_C, u_C) &\leq (\tilde{\mathbf{V}}_C^T \mathbf{K} \tilde{\mathbf{V}}_C u_C, u_C) \leq c_E^2 (\mathbf{S}_C u_C, u_C) \\ \forall u_C &\in \mathbb{R}^{N_C} \end{aligned} \quad (113)$$

which are the matrix form of the inequalities

$$\begin{aligned} a(P^* u_C, P^* u_C) &\leq a(\tilde{P} u_C, \tilde{P} u_C) \\ &\leq c_E^2 a(P^* u_C, P^* u_C) \quad \forall u_C \in \mathbf{V}|_{\Gamma_C} \end{aligned} \quad (114)$$

Of course, replacing P^* by \tilde{P} , we lose the orthogonality. However, the following lemma shows that the bounded extension gives us a stable splitting.

Lemma 2. If (108) holds, then

$$\gamma = \cos \angle(\tilde{\mathbf{V}}_C, \mathbf{V}_I) \leq \sqrt{1 - c_E^{-2}} < 1 \quad (115)$$

Proof. Follows immediately from relation (62) in Lemma 1 (see e.g. Haase *et al.*, 1994). \square

The sharp (minimal) constant c_E in (112)–(113) providing also the sharp constant in (115) is given by the maximal eigenvalue λ_{\max} of the generalized eigenvalue problem $\tilde{\mathbf{V}}_C^T \mathbf{K} \tilde{\mathbf{V}}_C u_C = \lambda \mathbf{S}_C u_C$. Now, we can summarize our results in the following theorem.

Theorem 8. Assume that there is some bounded extension $\tilde{\mathbf{P}}_{IC} : \mathbb{R}_{N_C} \rightarrow \mathbb{R}_{N_I}$ satisfying (112) with some constant $c_E \geq 1$ and that there are SPD preconditioners \mathbf{C}_C and \mathbf{C}_I for $\tilde{\mathbf{V}}_C^T \mathbf{K} \tilde{\mathbf{V}}_C$ and \mathbf{K}_I respectively, that is, there are positive spectral equivalence constants γ_C, γ_I , and $\bar{\gamma}_I$ such that

$$\begin{aligned} \gamma_C \mathbf{C}_C &\leq \tilde{\mathbf{V}}_C^T \mathbf{K} \tilde{\mathbf{V}}_C \leq \bar{\gamma}_C \mathbf{C}_C \\ \text{and} \quad \gamma_I \mathbf{C}_I &\leq \mathbf{K}_I \leq \bar{\gamma}_I \mathbf{C}_I \end{aligned} \quad (116)$$

Then, the inexact ASM preconditioner

$$\mathbf{C} = \begin{pmatrix} \mathbf{I}_C & -\tilde{\mathbf{P}}_{IC}^T \\ \mathbf{0} & \mathbf{I}_I \end{pmatrix} \begin{pmatrix} \mathbf{C}_C & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_I \end{pmatrix} \begin{pmatrix} \mathbf{I}_C & \mathbf{0} \\ -\tilde{\mathbf{P}}_{IC} & \mathbf{I}_I \end{pmatrix} \quad (117)$$

based on the bounded extension splitting (105) is spectrally equivalent to \mathbf{K} , that is,

$$\underline{\gamma} \mathbf{C} \leq \mathbf{K} \leq \bar{\gamma} \mathbf{C} \quad (118)$$

with the spectral equivalence constants

$$\underline{\gamma} = \min\{\underline{\gamma}_C, \underline{\gamma}_I\} \left(1 - \sqrt{1 - c_E^2}\right) \quad (119)$$

and

$$\bar{\gamma} = \max\{\bar{\gamma}_C, \bar{\gamma}_I\} \left(1 + \sqrt{1 - c_E^2}\right)$$

Proof. Follows easily from Corollary 1 and Lemma 2. \square

Theorem 8 provides us with a guide for choosing the ingredients \mathbf{C}_I , \mathbf{C}_C and \mathbf{E}_C of the inexact ASM preconditioner (116) in such a way that the final spectral equivalence constants $\underline{\gamma}$ and $\bar{\gamma}$ in (113) do not depend on h , H and the jumps of coefficient too much. Optimal ingredients will lead to an optimal inexact ASM preconditioner. Let us summarize some concrete proposals for choosing \mathbf{C}_I , \mathbf{C}_C , and \mathbf{E}_C :

Preconditioners \mathbf{C}_I for the local Dirichlet problems

Since $\mathbf{C}_I = \text{diag}(\mathbf{C}_{I_j})_{j=1,\dots,J}$, we only need good preconditioners \mathbf{C}_{I_j} for the \mathbf{K}_{I_j} , arising from the FE discretization of the local Dirichlet problems where the coefficients of the PDE are changing only smoothly. Nowadays, a lot of optimal (linear complexity) preconditioners for such problems are available (see e.g. Chapter 20, this Volume or Bramble and Zhang, 2000). For instance, local multigrid preconditioners of the form

$$\mathbf{C}_{I_j} = \mathbf{K}_{I_j} (\mathbf{I}_j - \mathbf{E}_{I_j}^k)^{-1} \quad (120)$$

will do a good job, where \mathbf{E}_{I_j} denotes the corresponding multigrid iteration operator. They can be generated, for example, by one ($k=1$) symmetric V-cycle with appropriately chosen multigrid components such that \mathbf{C}_{I_j} is SPD (Jung and Langer, 1991). Since we can assume that the multigrid iteration operators \mathbf{E}_{I_j} are nonnegative with respect to the energy inner product and the multigrid rates $\|\mathbf{E}_{I_j}\|_{\mathbf{K}_{I_j}}$ are bounded by some mesh-independent constant $\eta < 1$, we see that the second spectral inequalities in (116) are fulfilled by $\underline{\gamma}_I = 1 - \eta^k$ and $\bar{\gamma}_I = 1$. The operation count for the local preconditioning operation gives $\text{ops}(\mathbf{C}_{I_j}^{-1} \mathbf{d}_{I_j}) = O(N_{I_j}) = O((H/h)^d)$. Whereas various optimal preconditioners are available in the h -version of the FEM, the situation is quite different for the hp -version. Examples of optimal, or at least almost optimal local hp preconditioners can be found in Korneev and Jensen (1999),

Korneev (2001, 2002a), Beuchler (2002), and Korneev, Langer and Xanthis (2003).

Preconditioners \mathbf{C}_C for $\tilde{\mathbf{V}}_C^T \mathbf{K} \tilde{\mathbf{V}}_C$
Owing to the spectral relations (113), every good Schur-complement preconditioner \mathbf{C}_C , given in Section 5.2, is also a good preconditioner for $\tilde{\mathbf{V}}_C^T \mathbf{K} \tilde{\mathbf{V}}_C$ provided that the extension constant c_E is small. More precisely, the spectral equivalence inequalities $\underline{s}_C \mathbf{C}_C \leq \mathbf{S}_C \leq \bar{s}_C \mathbf{C}_C$ and (109) give us the first spectral inequalities in (116) with $\underline{\gamma}_C = \underline{s}_C$ and $\bar{\gamma}_C = \bar{s}_C \bar{s}_C$. On the other hand, one can again construct preconditioners of the form

$$\mathbf{C}_C = \tilde{\mathbf{V}}_C^T \mathbf{K} \tilde{\mathbf{V}}_C (\mathbf{I}_C - \mathbf{E}_C^k)^{-1} \quad (121)$$

applying s iteration steps of some symmetric internal iteration method (e.g. s symmetric V-cycles) with the linear iteration operator \mathbf{E}_C directly to the ASM subspace matrix $\tilde{\mathbf{V}}_C^T \mathbf{K} \tilde{\mathbf{V}}_C$. Mention that the discrete bounded extension operations discussed below are cheap such that in turn the cost for the matrix-by-vector multiplication $\tilde{\mathbf{V}}_C^T \mathbf{K} \tilde{\mathbf{V}}_C \mathbf{d}_C$ is proportional to N .

Discrete bounded extension
Owing to the equivalence

$$\begin{aligned} \frac{\theta}{2} \|\mathbf{u}\|_{\mathbf{H}^{1/2}(\Gamma_j)}^2 &\leq \inf_{\mathbf{v} \in \mathbf{H}^1(\Omega_j): \mathbf{v}|_{\Gamma_j} = \mathbf{u}|_{\Gamma_j}} \|\nabla \mathbf{v}\|_{\mathbf{L}_2(\Omega_j)}^2 \\ &\leq \bar{\theta} \|\mathbf{u}\|_{\mathbf{H}^{1/2}(\Gamma_j)}^2 \quad \forall \mathbf{u} \in \mathbf{H}^{1/2}(\Gamma_j) \end{aligned} \quad (122)$$

and to (102), we can reduce the construction of a bounded extension operator $\tilde{\mathbf{P}}_j$ to the construction of a local bounded extension operator $\tilde{\mathbf{P}}_j : \mathbf{V}|_{\Gamma_j} \subset \mathbf{H}^{1/2}(\Gamma_j) \rightarrow \mathbf{V}|_{\Omega_j} \subset \mathbf{H}^1(\Omega_j)$ such that the inequality

$$\|\tilde{\mathbf{P}}_j \mathbf{u}\|_{\mathbf{H}^1(\Omega_j)} = \|\nabla \tilde{\mathbf{P}}_j \mathbf{u}\|_{\mathbf{L}_2(\Omega_j)} \leq \tilde{c}_E \|\mathbf{u}\|_{\mathbf{H}^{1/2}(\Gamma_j)} \quad \forall \mathbf{u} \in \mathbf{V}|_{\Gamma_j} \quad (123)$$

is valid for some positive constant \tilde{c}_E . If \tilde{c}_E is independent of H/h , then the extension constant $c_E^2 = \tilde{c}_E^2/\theta$ does not depend on these parameters either.

Let us now review some cheap bounded discrete extension procedures of that kind. The first computable extension procedure was proposed by Matsokin and Nepomnyaschikh (1985) on the basis of some averaging technique that provides a uniform bound (see also Nepomnyaschikh, 1991a). Haase *et al.* (1994) introduced the hierarchical extension that is very cheap. In 2D, it leads to a $\ln(H/h)$ growth of c_E that can be compensated by $O(\ln(H/h))$ multigrid iterations. In 3D, the hierarchical extension is too weak. However, the multilevel extension that was proposed by Haase and Nepomnyaschikh (1997) works fine in 2D as well as in 3D.

5.4 Neumann–Neumann preconditioners

Bourgat *et al.* (1989) introduced the Neumann–Neumann Schur-complement preconditioner (cf. also Section 5.2) $\mathbf{C}_C^{-1} = (1/4)\mathbf{S}_C^{-1} + (1/4)\mathbf{S}_C^{-1}$ for the case of two subdomains ($J=2$) and showed that $\kappa(\mathbf{C}_C^{-1}\mathbf{S}_C) = O(1)$. The operation $\mathbf{w}_C = \mathbf{S}_C^{-1} \mathbf{d}_C$ ($\mathbf{d}_{C_1} = \mathbf{d}_{C_2} = \mathbf{d}_C$) is obviously equivalent to solution of the system

$$\begin{pmatrix} \mathbf{K}_{C_1} & \mathbf{K}_{C_1 C_2} \\ \mathbf{K}_{C_2 C_1} & \mathbf{K}_{C_2} \end{pmatrix} \begin{pmatrix} \mathbf{w}_{C_1} \\ \mathbf{w}_{C_2} \end{pmatrix} = \begin{pmatrix} \mathbf{d}_{C_1} \\ \mathbf{d}_{C_2} \end{pmatrix} \quad (124)$$

that corresponds to the Neumann boundary condition on $\Gamma_C = \partial\Omega_1 \cap \partial\Omega_2$ for $j=1, 2$.

De Roeck and Le Tallec (1991) generalized the Neumann–Neumann preconditioner to the general case of J subdomains. To simplify the notation, we skip the subindex C , that is, $\mathbf{S}_j = \mathbf{S}_{C_j}$, $\mathbf{A}_j = \mathbf{A}_{C_j}$, and so on. In order to weight the contributions from the different subdomains Ω_j , we introduce the weight matrices \mathbf{D}_j such that $\sum_{j=1}^J \mathbf{A}_j^T \mathbf{D}_j \mathbf{A}_j = \mathbf{I}_J$. There are different ways to choose these weights. Following Mandel and Brezina (1996), we define \mathbf{D}_j as the diagonal matrix $\text{diag}(\mathbf{d}_j^i)_{i=1, \dots, N_j}$ with the diagonal entries

$$d_j^i = \alpha_j / \sum_{i: x_i \in \partial\Omega_j} \alpha_i \quad (125)$$

This choice avoids the dependence of the condition number on the coefficient jumps in the PDE. In the case that all $\alpha_j = 1$ (Poisson equation), the diagonal entry d_j^i is equal to the reciprocal of the number of subdomains meeting at the nodal point x_i to which the diagonal entry d_j^i belongs. Similar to the case of two subdomains, we can now write the multisubdomain Neumann–Neumann preconditioner \mathbf{C}

in the form

$$\mathbf{C}^{-1} = \sum_{j=1}^J \mathbf{A}_j^T \mathbf{D}_j^T \mathbf{S}_j^{-1} \mathbf{D}_j \mathbf{A}_j \quad (126)$$

that is again an additive Schwarz preconditioner. The algorithmic form of the preconditioning operation $\mathbf{w} = \mathbf{C}^{-1} \mathbf{d}$ is given in Algorithm 7. In general ($\partial\Omega_j \cap \Gamma_D = \emptyset$), the local Schur complements \mathbf{S}_j as well as the corresponding subdomain stiffness matrices in (124) are singular because they are derived from pure local Neumann problems. Thus, the kernels correspond to the functions that are constant in the subdomain Ω_j (Example 1) and to local rigid body motions in the case of linear elasticity (Example 2). Therefore, to ensure solvability, the right-hand sides of the systems must be orthogonal to the kernel (this is in general not the case!), and if they are orthogonal to the kernel, then the solution is not unique. Another serious drawback of Algorithm 7 consists in the absence of some global information exchange mechanism that causes the H^{-2} dependence of the relative condition number $\kappa(\mathbf{C}^{-1}\mathbf{S})$. The balancing technique introduced by Mandel (1993) removes both drawbacks (see also Dryja and Widlund (1995) for a different approach). Let us introduce $N_j \times M_j$ matrices \mathbf{Z}_j consisting of M_j linear independent column vectors $\mathbf{z}_j^m \in \mathbb{R}^{N_j}$ ($m=1, \dots, M_j$) such that

$$\ker \mathbf{S}_j \subset \text{range } \mathbf{Z}_j = \text{span}\{\mathbf{z}_j^1, \dots, \mathbf{z}_j^{M_j}\} \quad (127)$$

For our model problem (Example 1), we can simply choose $\ker \mathbf{S}_j = \text{range } \mathbf{Z}_j = \text{span}\{(1, \dots, 1)^T\}$ for the singular case and omit the balancing procedures (128) and (130) in Algorithm 8 for those i 's which belong to the regular local Schur complements \mathbf{S}_j . Then it is clear that some vector \mathbf{r} fulfills the local orthogonality conditions ensuring the solvability of the local Neumann problems if $\mathbf{Z}_j^T \mathbf{D}_j \mathbf{A}_j \mathbf{d} = 0$ for the

Algorithm 7. Neumann–Neumann preconditioning operation $\mathbf{w} = \mathbf{C}^{-1} \mathbf{d}$.

```

 $\mathbf{d} \in \mathbb{R}^{N_C}$  given vector (defect/residual) [initialization]
for all  $j = 1, \dots, J$  do
   $\mathbf{d}_j = \mathbf{D}_j \mathbf{A}_j \mathbf{d}$  (distribute  $\mathbf{d}$  to the subdomains)
end for
for all  $j \in \{1, \dots, J\}$  in parallel do {solve the local Neumann problems (124) in parallel}
   $\mathbf{S}_j \mathbf{w}_j = \mathbf{d}_j$ 
end for
 $\mathbf{w} = \sum_{j=1}^J \mathbf{A}_j^T \mathbf{D}_j^T \mathbf{w}_j$  (average the local solution on the interface)
```

J 's corresponding to the singular cases. Adding this balancing step in a symmetric way to Algorithm 7, we arrive at the so-called balancing Neumann-Neumann preconditioning Algorithm 8.

The following remarks may be useful for the practical implementation of Algorithm 8:

1. On the one hand, the balancing steps (128) and (130) can be omitted for such subdomain Ω_i where the subdomain problems are regular, that is, $M_i = 0$, no λ_i and μ_i . On the other hand, the spaces $\text{range } Z_i$ can always be enriched.
2. The sparse SPD matrix arising from the auxiliary problems (128) and (130) has the dimension $M \times M$ with $M = \sum_{j=1}^J M_j$. This matrix can easily be generated in a preprocessing step.
3. Owing to the postbalancing step (130), any solution of the local Neumann problems (129) will be appropriate.

4. If the old residual \mathbf{d} is already balanced, then the prebalancing step (128) can be omitted, that is, if the initial residual of the Schur-complement CG iteration is balanced, then the prebalancing step (128) can always be omitted.

From Algorithm 8, we observe that the balanced Neumann-Neumann preconditioner can be rewritten in the compact form $\mathbf{C}^{-1} = (\mathbf{I} - \mathbf{E})\mathbf{S}^{-1}$ with the iteration matrix

$$\mathbf{E} = (\mathbf{I} - \mathbf{P})(\mathbf{I} - \mathbf{TS})(\mathbf{I} - \mathbf{P}) \\ = (\mathbf{I} - \mathbf{P}) \left[\mathbf{I} - \left(\sum_{j=1}^J \mathbf{A}_j^T \mathbf{D}_j^T \mathbf{S}_j^+ \mathbf{D}_j \mathbf{A}_j \right) \mathbf{S} \right] (\mathbf{I} - \mathbf{P}) \quad (131)$$

where \mathbf{S}_j^+ denotes the Moore-Penrose pseudo-inverse of \mathbf{S}_j , and \mathbf{P} is the \mathbf{S} -orthogonal projection onto the space $\mathbf{P} = \{ \mathbf{v} = \sum_{j=1}^J \mathbf{A}_j^T \mathbf{D}_j^T \mathbf{z}_j \in \mathbb{R}^N : \mathbf{z}_j \in \text{range } \mathbf{Z}_j \}$ that plays the role of some 'coarse grid space'. Thus, the balancing

Algorithm 8. Balancing Neumann-Neumann preconditioning operation $\mathbf{w} = \mathbf{C}^{-1}\mathbf{d}$.

$\mathbf{d} \in \mathbb{R}^N$ given vector (defect/residual) {initialization}
Find $\lambda_i \in \mathbb{R}^{M_i}$, $i = 1, \dots, J$, such that [balancing the old residual vector]

$$\mathbf{Z}_i^T \mathbf{D}_i \mathbf{A}_i \left(\mathbf{d} - \mathbf{S} \sum_{j=1}^J \mathbf{A}_j^T \mathbf{D}_j^T \mathbf{z}_j \right) = 0 \quad \forall i = 1, \dots, J \quad (128)$$

and set $\mathbf{r} = \mathbf{d} - \mathbf{S} \sum_{j=1}^J \mathbf{A}_j^T \mathbf{D}_j^T \mathbf{z}_j$

for all $j = 1, \dots, J$ do

$\mathbf{r}_j = \mathbf{D}_j \mathbf{A}_j \mathbf{r}$

end for

(distribute \mathbf{r} to the subdomains)

(solve the balanced local Neumann problems (124) in parallel)

for all $j \in \{1, \dots, J\}$ in parallel do

$$\mathbf{S}_j \mathbf{w}_j = \mathbf{r}_j \quad (129)$$

end for

Find $\mu_i \in \mathbb{R}^{M_i}$, $i = 1, \dots, J$, such that

(balancing the new residual vector)

$$\mathbf{Z}_i^T \mathbf{D}_i \mathbf{A}_i \left(\mathbf{d} - \mathbf{S} \sum_{j=1}^J \mathbf{A}_j^T \mathbf{D}_j^T (\mathbf{w}_j + \mathbf{Z}_j \mu_j) \right) = 0 \quad \forall i = 1, \dots, J \quad (130)$$

$$\mathbf{w} = \sum_{j=1}^J \mathbf{A}_j^T \mathbf{D}_j^T (\mathbf{w}_j + \mathbf{Z}_j \mu_j)$$

(average the local solution on the interface)

steps are handling not only the singular local Neumann problems but also the global information transport via the space \mathbf{P} . This is the reason why some enrichment of space can be meaningful. From (131), we immediately see that the balanced Neumann-Neumann preconditioner is some kind of hybrid Schwarz preconditioner, as described in Section 3.2.3.

Theorem 9. The balanced Neumann-Neumann preconditioner $\mathbf{C} = \mathbf{S}(\mathbf{I} - \mathbf{E})^{-1}$ defined by Algorithm 8 is SPD and

$$\kappa(\mathbf{C}^{-1}\mathbf{S}) \leq \sup_{\mathbf{u}_j \in \mathbf{X}_j, j=1, \dots, J} \frac{\sum_{i=1}^J \|\mathbf{A}_i\| \sum_{j=1}^J \|\mathbf{A}_j^T \mathbf{D}_j^T \mathbf{u}_j\|_{\mathbf{S}_i}^2}{\sum_{j=1}^J \|\mathbf{u}_j\|_{\mathbf{S}_j}^2} \\ \leq c \left(1 + \log \frac{H}{h} \right)^2 \quad (132)$$

where $\mathbf{X}_j = \{ \mathbf{u}_j \in \mathbb{R}^{N_j} : (\mathbf{u}_j, \mathbf{v}_j) = 0 \, \forall \mathbf{v}_j \in \ker \mathbf{S}_j \text{ and } (\mathbf{S}_j \mathbf{u}_j, \mathbf{v}_j) = 0 \, \forall \mathbf{v}_j \in \text{range } \mathbf{Z}_j \}$, and c denotes a positive constant that is independent of h , H , and the jumps in the coefficients.

Proof. The first part of estimate (132) was proved by Mandel (1993). This estimate is based on pure linear algebra arguments and is not directly connected to our model problem. This abstract estimate was used by Mandel and Brezina (1996) to produce the bound at the right-hand side of estimate (132) for our model problem. \square

The advantages of the balanced Neumann-Neumann Schur-complement preconditioners are certainly their (almost) independence of bad parameters (see also Mandel and Brezina (1996) for impressive numerical results) and the fact that more or less standard software routines can be used. On the other hand, the balanced Neumann-Neumann preconditioned Schur-complement CG that is mostly used in practice is quite expensive with respect to the number of arithmetical operations because one Dirichlet and one Neumann problem must be solved exactly (directly) per subdomain (however, completely in parallel) and per iteration step. Inexact versions are not straightforward (see, however, Sections 3 and 5.3).

5.5 Finite element tearing and interconnecting methods

The FETI methods were introduced by Farhat and Roux (1991) (see also Farhat and Roux (1994) for a more detailed description by the same authors) as a nonoverlapping DD parallel solution method for our system (13)

of conform finite element equations that can be reduced to the Schur-complement system (95) after eliminating the internal unknowns, as described in Section 5.1. Tearing the unknowns \mathbf{u}_Γ at the interface Γ_C first into independent unknowns building the vectors $\mathbf{u}_1, \dots, \mathbf{u}_J$ (from now on we again omit the index C , as in the preceding section) and then again enforcing the continuity by simple interconnecting constraints $\mathbf{B}\mathbf{u} = \mathbf{0}$, we arrive at the saddle-point problem: Given $\mathbf{f} = (\mathbf{f}_j)_{j=1, \dots, J} \in \mathbb{R}^N$, find $\mathbf{u} = (\mathbf{u}_j)_{j=1, \dots, J} \in \mathbf{U} = \mathbf{U}_1 \times \dots \times \mathbf{U}_J = \mathbb{R}^N$ and $\lambda \in \Lambda = \text{range } \mathbf{B} \subset \mathbb{R}^M$ such that

$$\begin{pmatrix} \mathbf{S} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{0} \end{pmatrix} \quad (133)$$

where $\mathbf{S} = \text{diag}(\mathbf{S}_j)_{j=1, \dots, J}$ denotes the $N \times N$ block diagonal matrix with the subdomain Schur complements on the diagonal, $\mathbf{B} = (\mathbf{B}_1, \dots, \mathbf{B}_J)$ is the $M \times N$ matrix of constraints that measures the jump of a given vector \mathbf{u} across the interface, $N = N_1 + \dots + N_J$, and M is the number of Lagrange multipliers. Each row of the matrix \mathbf{B} is connected with a pair of matching nodes across the interface. The entries of such a row are 1, -1, and 0 for the indices corresponding to the matching nodes and otherwise. Therefore, $\mathbf{B}\mathbf{u} = \mathbf{0}$ implies that the finite element function \mathbf{u} corresponding to \mathbf{u} is continuous across the interface Γ_C . We assume here that the number of constraints at some matching node is equal to the number of matching subdomain minus 1. This method of the minimal number of constraints resp. multipliers is called nonredundant (see e.g. Klawonn and Widlund (2001) for the use of redundant constraints).

Since $\ker \mathbf{S} \cap \ker \mathbf{B} = \{\mathbf{0}\}$, the saddle-point system (133) has a unique solution and is completely equivalent to the Schur-complement problem (95). The subdomain Schur complement \mathbf{S}_j is singular if the corresponding subdomain Ω_j does not touch the Dirichlet boundary Γ_D , that is $\Gamma = \partial\Omega$ for our model problem. Such subdomains are called floating subdomains. Similar to Section 5.4, we assume that $\ker \mathbf{S}$ can be represented by the range of some $N \times L$ matrix \mathbf{Z} , that is, now $\ker \mathbf{S} = \text{range } \mathbf{Z}$, with L being the number of floating subdomains. If we assume for the time being that the solvability condition

$$\mathbf{f} - \mathbf{B}^T \lambda \perp \ker \mathbf{S} = \text{range } \mathbf{Z}, \text{ i.e. } \mathbf{Z}^T (\mathbf{f} - \mathbf{B}^T \lambda) = \mathbf{0} \quad (134)$$

for first equation in (133) is fulfilled, then the solution \mathbf{u} can be represented in the form

$$\mathbf{u} = \mathbf{S}^+ (\mathbf{f} - \mathbf{B}^T \lambda) + \mathbf{Z} \alpha \quad (135)$$

with some element $\mathbf{Z} \alpha \in \ker \mathbf{S}$ that has to be determined. Substituting now (135) into the second block equation of

(133), we arrive at the dual problem

$$\mathbf{B}\mathbf{S}^T\mathbf{B}^T\lambda = \mathbf{B}\mathbf{S}^T\mathbf{f} + \mathbf{G}\alpha \quad (136)$$

for defining λ and α with the abbreviation $\mathbf{G} = \mathbf{B}\mathbf{Z}$. Defining now the orthogonal projection $\mathbf{P} = \mathbf{I} - \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$ from the space Λ onto the subspace $\Lambda_0 = \ker \mathbf{G}^T = (\text{range } \mathbf{G})^\perp$ with respect to the scalar product $(\cdot, \cdot) = (\cdot, \cdot)_\Lambda = (\cdot, \cdot)_{\mathbf{H}^1(\Omega)}$, we can split the definition of λ from the definition of α . Indeed, applying \mathbf{P} to (136) gives the equation

$$\mathbf{P}\mathbf{B}\mathbf{S}^T\mathbf{B}^T\lambda = \mathbf{P}\mathbf{B}\mathbf{S}^T\mathbf{f} \quad (137)$$

since $\mathbf{P}\mathbf{G}\alpha = \mathbf{0}$. Together with the solvability condition (134), we get the final dual problem in the following form: Find $\lambda \in \Lambda$ such that

$$\mathbf{P}\mathbf{F}\lambda = \mathbf{P}\mathbf{d} \text{ subject to } \mathbf{G}^T\lambda = \mathbf{e} \quad (138)$$

with the abbreviations $\mathbf{F} = \mathbf{B}\mathbf{S}^T\mathbf{B}^T$, $\mathbf{d} = \mathbf{B}\mathbf{S}^T\mathbf{f}$, and $\mathbf{e} = \mathbf{Z}^T\mathbf{f}$. Once λ is defined, from (136) we obtain

$$\alpha = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T(\mathbf{F}\lambda - \mathbf{d}) \quad (139)$$

and, finally, we get \mathbf{u} from (135). The solution \mathbf{u}_C of the Schur-complement problem (95) can easily be extracted from \mathbf{u} .

The dual problem (138) is now solved by the preconditioned conjugate gradient (PCG) iteration in the subspace Λ_0 that is presented in Algorithm 9 as a projected PCG method.

The matrix-by-vector multiplication $\mathbf{F}\mathbf{s}^n = \mathbf{B}\mathbf{S}^T\mathbf{B}^T\mathbf{s}^n$ means the concurrent (direct) solution of J local Neumann problems. The orthoprojection \mathbf{P} ensures the solvability of the Neumann problems and the global information exchange. The application of $\mathbf{P} = \mathbf{I} - \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$ to some vector $\mathbf{w} \in \Lambda$ involves the direct solution of a small system with the $L \times L$ system matrix $\mathbf{G}^T\mathbf{G}$ that plays the role of some kind of a coarse grid problem. The FETI preconditioner \mathbf{C} should be spectrally equivalent to the FETI operator \mathbf{F} on the subspace $\Lambda_0 = \ker \mathbf{G}^T$, that is,

$$\gamma(\mathbf{C}\lambda, \lambda) \leq (\mathbf{F}\lambda, \lambda) \leq \bar{\gamma}(\mathbf{C}\lambda, \lambda) \quad \forall \lambda \in \Lambda_0 \quad (140)$$

with positive spectral equivalence constants γ and $\bar{\gamma}$ such that $\kappa(\mathbf{P}\mathbf{C}^{-1}\mathbf{P}^T\mathbf{F}\mathbf{P}) \leq \gamma/\bar{\gamma}$ is as small as possible and the preconditioning operation $\mathbf{C}^{-1}\mathbf{d}$ is as cheap as possible. Farhat and Roux (1991) proposed the FETI preconditioner

$$\mathbf{C}^{-1} = \mathbf{B}\mathbf{S}\mathbf{B}^T \quad (141)$$

Algorithm 9. FETI subspace PCG iteration.

```

{initialization}
 $\lambda^0 = \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{e}$  {forcing the constraints  $\mathbf{G}^T\lambda^0 = \mathbf{e}$  for the initial guess}
 $\mathbf{d}^0 = \mathbf{P}(\mathbf{d} - \mathbf{F}\lambda^0)$  {compute the defect and project to the subspace  $\Lambda_0$ }
 $\mathbf{w}^0 = \mathbf{C}^{-1}\mathbf{d}^0$  {precondition step}
 $\mathbf{s}^0 = \mathbf{z}^0 = \mathbf{P}\mathbf{w}^0$  {project the correction to the subspace  $\Lambda_0$ }
 $\beta_0 = (\mathbf{w}^0, \mathbf{d}^0) = (\mathbf{z}^0, \mathbf{d}^0)$  {begin iteration loop}

for  $n = 0$  step 1 until  $\beta_n \leq \varepsilon^2\beta_0$  do
   $\mathbf{x}^n = \mathbf{P}\mathbf{F}\mathbf{s}^{n-1}$  {matrix-by-vector multiplication + projection}
   $\alpha_n = (\mathbf{x}^n, \mathbf{s}^{n-1})$ 
   $\alpha = \beta_n/\alpha_n$ 
   $\lambda^{n+1} = \lambda^n + \alpha \mathbf{s}^{n-1}$  {update of the iterate in the subspace  $\Lambda_0$ }
   $\mathbf{d}^{n+1} = \mathbf{d}^n - \alpha \mathbf{x}^n$  {update of the defect in the subspace  $\Lambda_0$ }
   $\mathbf{w}^{n+1} = \mathbf{C}^{-1}\mathbf{d}^{n+1}$  {precondition step}
   $\mathbf{z}^{n+1} = \mathbf{P}\mathbf{w}^{n+1}$  {project the correction to the subspace  $\Lambda_0$ }
   $\beta_{n+1} = (\mathbf{w}^{n+1}, \mathbf{d}^{n+1}) = (\mathbf{z}^{n+1}, \mathbf{d}^{n+1})$ 
   $\beta = \beta_{n+1}/\beta_n$ 
   $\mathbf{s}^{n+1} = \mathbf{z}^{n+1} - \beta \mathbf{s}^n$  {update of the search direction in the subspace  $\Lambda_0$ }
end for

{end iteration loop}
```

that is now called the Dirichlet preconditioner because the multiplication of \mathbf{S} with some vector requires the concurrent solution of J Dirichlet problems. Mandel and Tezaur (1996) proved that the relative spectral condition number is rigorously bounded by $c(1 + \log(H/h))^2$. This polylogarithmic bound could be improved to $c(1 + \log(H/h))^2$ for special domain decompositions that do not contain cross points. Numerical studies with the Dirichlet FETI preconditioner (141) can be found in Farhat and Roux (1991) and Stefanica (2001). Similar to the balanced Neumann-Neumann Schur-complement preconditioning technique, Klawonn and Widlund (2001) introduced the preconditioner

$$\mathbf{C}^{-1} = (\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T)^{-1}\mathbf{B}\mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}\mathbf{B}^T(\mathbf{B}^T\mathbf{D}^{-1}\mathbf{B})^{-1} \quad (142)$$

with the diagonal scaling matrix $\mathbf{D} = \text{diag}(\mathbf{D}_j)_{j=1,\dots,J}$. This scaling and the introduction of an appropriately scaled scalar product in Λ_0 (this affects the orthogonal projection \mathbf{P}) lead to the rigorous bound $c(1 + \log(H/h))^2$ where the constant c is now independent of the jumps in the coefficients. The numerical and the parallel performances of the classical Dirichlet preconditioner and this new preconditioner are compared in Stefanica (2001). In its exact version, the FETI Algorithm 9 requires the exact (direct) solution of one local Neumann problem (matrix multiplication step) per subdomain (i.e. in parallel) at each iteration step. In these parts, local direct solvers can be very expensive in practical applications where very complex local problems can appear. The use of inexact solvers (preconditioners) for the local Dirichlet problems is more or less straightforward, whereas the replacement of exact Neumann solvers by inexact ones is not.

Klawonn and Widlund (2000) avoided the reduction of the original problem (13) to the saddle-point problem (133) by eliminating the internal unknowns and related the original problem (13) directly to the saddle-point problem

$$\begin{pmatrix} \mathbf{K} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{0} \end{pmatrix} \quad (143)$$

with the torn original stiffness matrix $\mathbf{K} = \text{diag}(\mathbf{K}_j)_{j=1,\dots,J}$ and the interconnecting matrix \mathbf{B} . We refer the reader to Farhat, Lesoinne and Pierson (2000) for some further development of the FETI methodology. In particular, the dual-primal FETI (FETI-DP) methods, introduced by Farhat *et al.* (2001), seem to be very attractive because they avoid the solution of singular problems on the subdomains by fixing some primal unknowns. The first results on the convergence analysis of FETI-DP methods were given by Mandel and Tezaur (2001) and Klawonn and Widlund

(2002). Langer and Steinbach (2003) introduced the boundary element tearing and interconnecting (BETI) methods as boundary element counterparts of the FETI methods.

5.6 Mortar methods

In the classical FETI method, we tore (split) the unknowns on the interface, which are conform in our original FE scheme, and interconnect them again by simple $(1, -1)$ equality constraints with the only aim to construct a DD solver that is essentially based on the dual problem for the corresponding Lagrange multipliers. The Mortar technique proposed by Bernardi and Maday and Patera (1993, 1994) goes one step further and allows the triangulation to be nonconforming. Thus, the FE solution cannot globally conform in this general situation and the continuity must be enforced by constraints in an appropriate way. This continuity constraints can be included into the product FE space or can be incorporated by Lagrange multipliers in a saddle-point formulation.

Let us again consider our model problem of Example 1 with piecewise constant coefficients and the nonoverlapping domain decomposition (92), but now we allow the triangulation and the FE functions to be nonconforming across the interfaces $\partial\Omega_j \cap \partial\Omega_i$. Thus, we look for a nonconforming FE solution \mathbf{u} in the product FE space $\mathbf{U} = \mathbf{V}_1 \times \dots \times \mathbf{V}_J$, where the subdomain FE spaces $\mathbf{V}_j = \mathbf{V}_h(\Omega_j) \subset \mathbf{H}^1(\Omega_j) \cap \mathbf{H}_0^1(\Omega)$ are defined on the individual subdomain triangulations \mathcal{T}_j using their individual finite elements. In order to get a proper approximation to our weak solution, we have to enforce weak continuity constraints by Lagrange multipliers in such a way that the approximation and consistency errors are not perturbed. To do this, we first introduce two different, but complementary nonoverlapping decompositions $\bar{\Gamma}_C = \bigcup_{j=1}^J \bigcup_{i \in \mathcal{M}(j)} \bar{\Gamma}_{ji}$ and $\bar{\Gamma}_C = \bigcup_{j=1}^J \bigcup_{i \in \mathcal{M}(j)} \bar{\Gamma}_{ij}$ of the interface Γ_C into mortar and nonmortar faces $\bar{\Gamma}_{ji} = \partial\Omega_j \cap \partial\Omega_i \subset \partial\Omega_j$ (edges in 2D). The face $\bar{\Gamma}_{ji}$ is considered as a part of $\partial\Omega_j$ and inherits the (surface) triangulation from Ω_j . If some face $\bar{\Gamma}_{ji}$ is mortar, that is, $i \in \mathcal{M}(j)$, then its opposite side $\bar{\Gamma}_{ij} \subset \partial\Omega_i$ is nonmortar, that is, $j \notin \mathcal{M}(i)$. Let us now introduce the discrete Lagrange multiplier space

$$\Lambda = \prod_{j=1}^J \prod_{i \in \mathcal{M}(j)} \Lambda(\bar{\Gamma}_{ji}) \subset \prod_{j=1}^J \prod_{i \in \mathcal{M}(j)} (\mathbb{H}^{1/2}(\bar{\Gamma}_{ji}))^* \quad (144)$$

where the local discrete Lagrange multiplier spaces are all connected with the nonmortar faces. The choice of local discrete Lagrange multiplier spaces $\Lambda(\bar{\Gamma}_{ji})$ is crucial not only for the approximation properties but also for efficiency reasons. For instance, in the case of linear triangular elements in 2D, the classical local discrete Lagrange multiplier space

$A(\Gamma_{ij})$ is a subspace of codimension two of the trace space $V_{h,\Gamma_{ij}}$ on the nonmortar edge Γ_{ij} . More precisely, $A(\Gamma_{ij})$ consists of all continuous, piecewise linear functions on Γ_{ij} , that are constant in the first and the last interval of the 1D mesh on Γ_{ij} induced by the mesh of $\bar{\Omega}_i$. We refer to Ben Belgacem and Maday (1999) for the 3D case and to Wohlmuth (2000) for biorthogonal mortar elements.

Now, the mortar scheme can be formulated in the constrained product space $V = \{v \in U : b(v, \mu) = 0 \forall \mu \in \Lambda\}$ as nonconforming DD FE scheme: Find $u \in V$ such that

$$a(u, v) = (f, v) \quad \forall v \in V \quad (145)$$

Alternatively, the mortar scheme can be reformulated as a mixed scheme in the unconstrained product space and the Lagrange multiplier space. Find $u \in U$ and $\lambda \in \Lambda$ such that

$$a(u, v) + b(v, \lambda) = (f, v) \quad \forall v \in U \quad (146)$$

$$b(u, \mu) = 0 \quad \forall \mu \in \Lambda \quad (147)$$

where $a(u, v) = \sum_{j=1}^L \alpha_j \int_{\Omega_j} \nabla u(x) \cdot \nabla v(x) dx$, $b(v, \mu) = \sum_{j=1}^L \sum_{i \in \mathcal{M}(j)} \int_{\Gamma_{ij}} [v] \mu ds$, and $[v] = v|_{\Gamma_{ij}^+} - v|_{\Gamma_{ij}^-}$ denotes the jump across face Γ_{ij} that geometrically coincides with Γ_{ij} .

The saddle-point problem can be rewritten in matrix form as the full FETI saddle-point problem (143), or, after eliminating the inner subdomain unknowns, as the reduced FETI saddle-point problem (133). However, the Lagrange multiplier matrix B is now defined by the mortar conditions (147) across the faces Γ_{ij} instead of the simple hard nodal continuity condition in the FETI method. Now, the FETI solver can be used for solving the mortar saddle-point problem in the same way as in Section 5.6 for solving the original FETI equations (133) (see Stefanica (2001) for more information and numerical experiments). We mention that other nonoverlapping DD algorithms and multilevel methods can successfully be applied to the solution of the mortar equation as well (see Wohlmuth (2001) for more information).

ACKNOWLEDGMENTS

The authors would like to thank the *Mathematisches Forschungsinstitut Oberwolfach* for the hospitality during their RiP stay in summer 2002, the *Austrian Science Fund (FWF)* for supporting the research work of the authors under the grant SFB F013, the Russian Fund of Basic Research, grant 00-01-00772, and the program "Universities of Russia", grant 03.01.017.

REFERENCES

- Ainsworth M. A preconditioner based on domain decomposition for h - p finite element approximation on quasi-uniform meshes. *SIAM J. Numer. Anal.* 1996; 33:1358–1376.
- Ainsworth M and Senior B. Aspects of an adaptive hp-finite element method: adaptive strategy, conforming approximation and efficient solvers. *Comput. Methods Appl. Mech. Eng.* 1997; 150:65–87.
- Aronszajn N. Theory of reproducing kernels. *Trans. Am. Math. Soc.* 1950; 68:337–404.
- Axelsson O and Vassilevskii PS. Algebraic multilevel preconditioning methods I. *SIAM J. Numer. Anal.* 1989; 56:157–177.
- Babuska I. On the Schwarz algorithm in the theory of differential equations of mathematical physics. *Czechoslov. Math. J.* 1958; 8(83):328–342 (in Russian).
- Babuska I, Craig A, Mandel J and Pitkäranta J. Efficient preconditioning for the p -version finite element method in two dimensions. *SIAM J. Numer. Anal.* 1991; 28(3):624–661.
- Bank RE, Jimack PK, Nadeem SA and Nepomnyashchikh SV. A weakly overlapping domain decomposition preconditioner for the finite element solution of elliptic partial differential equations. *SIAM J. Sci. Comput.* 2002; 23(6):1818–1842.
- Ben Belgacem F and Maday Y. The mortar element method for three dimensional finite elements. *Modell. Math. Anal. Numer.* 1999; 36(4):1234–1263.
- Bernardi C, Dauge M and Maday Y. Relèvements de traces préservant les polynômes. *C. R. Acad. Sci. Paris Sér. I Math.* 1992; 315:333–338 (in French).
- Bernardi C, Maday Y and Patera A. Domain decomposition by the mortar element method. In *Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters*, Kaper H and Garbey M (eds). Reidel: Dordrecht, 1993; 269–286.
- Bernardi C, Maday Y and Patera A. A new nonconforming approach to domain decomposition: the mortar element method. In *Nonlinear Partial Differential Equations and their Applications*, Brezis H and Lions JL (eds). Pitman: New York, 1994; 13–51.
- Bernardi C, Maday Y and Sacchi-Landriani G. Non conforming matching conditions for coupling spectral and finite element method. *Appl. Numer. Math.* 1989; 54:64–84.
- Betchler S. Multigrid solver for the inner problem in domain decomposition methods for p -FEM. *SIAM J. Numer. Anal.* 2002; 40(4):928–944.
- Björstad PE and Mandel J. On the spectra of sums of orthogonal projections with applications to parallel computing. *BIT* 1991; 31:76–88.
- Björstad PE and Widlund OB. Iterative methods for the solution of elliptic problems on regions partitioned into substructures. *SIAM J. Numer. Anal.* 1986; 23(6):1093–1120.
- Bourgat JP, Glowinski R, Le Tallec P and Vidrascu M. Variational formulation and algorithm for trace operator in domain decomposition calculations. In *Second International Symposium on Domain Decomposition Methods for Partial Differential Equations*, Chan TF, Glowinski R, Périaux J and Widlund OB (eds). SIAM: Philadelphia, 1989; 3–16.
- Bramble JH and Zhang X. The analysis of multigrid methods. *Handbook of Numerical Analysis VII*. North Holland: Amsterdam, 2000; 173–415.
- Bramble JH, Pasciak JE and Schatz AH. The construction of preconditioners for elliptic problems by substructuring, I. *Math. Comput.* 1986; 47(175):103–134.
- Bramble JH, Pasciak JE and Schatz AH. The construction of preconditioners for elliptic problems by substructuring, II. *Math. Comput.* 1987; 49(179):1–16.
- Bramble JH, Pasciak JE and Schatz AH. The construction of preconditioners for elliptic problems by substructuring, III. *Math. Comput.* 1988; 51(184):415–430.
- Bramble JH, Pasciak JE and Schatz AH. The construction of preconditioners for elliptic problems by substructuring, IV. *Math. Comput.* 1989; 53(187):1–24.
- Bramble JH, Pasciak JE and Xu J. Parallel multilevel preconditioners. *Math. Comput.* 1990; 55:1–22.
- Bramble JH, Pasciak JE, Wang J and Xu J. Convergence estimates for product iterative methods with applications to domain decomposition. *Math. Comput.* 1991; 57(195):1–21.
- Brenner SC. The condition number of the Schur complement in domain decomposition. *Numer. Math.* 1999; 83:187–203.
- Brenner SC. Lower bounds for two-level additive Schwarz preconditioners with small overlap. *SIAM J. Sci. Comput.* 2000; 21(5):1657–1669.
- Casarin MA. Quasi-optimal Schwarz methods for the conforming spectral element discretization. *SIAM J. Numer. Anal.* 1997; 34(6):2482–2502.
- Carstensen C, Kuhn M and Langer U. Fast parallel solvers for symmetric boundary element domain decomposition equations. *Numer. Math.* 1998; 79(2):321–347.
- Chan TF. Iterative methods for the solution of elliptic problems on regions partitioned into substructures. *SIAM J. Numer. Anal.* 1987; 24(2):382–390.
- Chan TF and Mathew TP. Domain decomposition algorithms. *Acta Numer.* 1994; 61–143.
- Ciarlet P. *The Finite Element Method for Elliptic Problems*. North Holland Publishing Company: Amsterdam, New York, Oxford, 1978.
- Demmel JW, Eisenstat TSC, Gilbert JR, Li XS and Liu JWH. A supernodal approach to sparse partial pivoting. *SIAM J. Matrix Anal. Appl.* 1999; 20:720–755.
- De Roeck YH and Le Tallec P. Analysis and test of local domain decomposition preconditioners. In *Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations*, Glowinski R, Kuznetsov YA, Meurant G, Périaux J and Widlund OB (eds). SIAM: Philadelphia, 1991; 112–128.
- Dryja M. A capacitance matrix method for Dirichlet problems on polygonal regions. *Numer. Math.* 1982; 39(1):51–64.
- Dryja M. A finite element-capacitance matrix method for elliptic problems on regions partitioned into substructures. *Numer. Math.* 1984; 44(1):153–168.
- Dryja M and Widlund OB. Some domain decomposition algorithm for elliptic problems. In *Iterative Methods for Large Linear Systems*, Hayes L and Kincaid D (eds). Academic Press: Orlando, 1989; 273–291.
- Dryja M and Widlund OB. Domain decomposition algorithms with small overlap. *SIAM J. Sci. Comput.* 1994; 15(3):604–620.
- Dryja M and Widlund OB. Schwarz methods of Neumann–Neumann type for three-dimensional elliptic finite element problems. *Commun. Pure Appl. Math.* 1995; 48(2):121–155.
- Dryja M, Smith BF and Widlund OB. Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions. *SIAM J. Numer. Anal.* 1994; 31(6):1662–1694.
- Farhat C and Roux FX. A method of finite element tearing and interconnecting and its parallel solution algorithm. *Int. J. Numer. Methods Eng.* 1991; 32:1205–1227.
- Farhat C and Roux FX. Implicit parallel processing in structural mechanics. In *Computational Mechanics Advances*, Oden JT (ed.). North Holland: Amsterdam, 1994; 1–124.
- Farhat C, Lesoinne M and Pierson K. A scalable dual-primal domain decomposition method. *Numer. Lin. Alg. Appl.* 2000; 7:687–714.
- Farhat C, Lesoinne M, Le Tallec P, Pierson K and Rixen D. FETI-DP: A dual-primal unified FETI method – part I: A faster alternative to the two-level FETI method. *Int. J. Numer. Methods Eng.* 2001; 50:1523–1544.
- George A and Liu JWH. *Computer Solution of Large Sparse Positive Definite Systems*. Prentice Hall Inc: Englewood Cliffs, 1981.
- Golub G and Mayers D. The use of preconditioning over irregular regions. In *Computing Methods in Applied Sciences and Engineering*, Glowinski R and Lions JL (eds). North Holland: Amsterdam, New York, Oxford, 1994; 3–14.
- Griebel M and Oswald P. On the abstract theory of additive and multiplicative Schwarz algorithms. *Numer. Math.* 1995; 70(2):163–180.
- Gupta A. Recent advances in direct methods for solving unsymmetric sparse systems of linear equations. *ACM Trans. Math. Softw.* 2002; 28:301–324.
- Haase G and Langer U. The non-overlapping domain decomposition multiplicative Schwarz method. *Int. J. Comput. Math.* 1992; 44:223–242.
- Haase G and Nepomnyashchikh SV. Extension explicite operators on hierarchical grids. *East-West J. Numer. Math.* 1997; 5(4):231–248.
- Haase G, Langer U and Meyer A. The approximate Dirichlet domain decomposition method. Part I and II. *Computing* 1991; 47:137–167.
- Haase G, Langer U, Meyer A and Nepomnyashchikh SV. Hierarchical extension operators and local multigrid methods in domain decomposition preconditioners. *East-West J. Numer. Math.* 1994; 2(3):173–193.
- Haase G, Heise B, Kuhn M and Langer U. Adaptive domain decomposition methods for finite and boundary element equations. In *Boundary Element Topics*, Wendland WL (ed.). Springer-Verlag: Berlin, Heidelberg, New York, 1997; 121–147.

- Ivanov SA and Korneev VG. Selection of the high order coordinate functions and preconditioning in the frame of the domain decomposition method. *Izv. Vysht. Uchebn. Zaved.* 1995; 39(4):62–81 (in Russian).
- Ivanov SA and Korneev VG. Preconditioning in the domain decomposition methods for the p -version with the hierarchical bases. *Math. Model.* 1996; 8(9):63–73.
- Jung M and Langer U. Application of multilevel methods to practical problems. *Surveys Math. Ind.* 1991; 1:217–237.
- Keyes DE and Gropp WD. A comparison of domain decomposition techniques for elliptic partial differential equations and their parallel implementation. *SIAM J. Sci. Comput.* 1987; 8(2):166–202.
- Klawonn A and Widlund OB. A domain decomposition method with Lagrange multipliers and inexact solvers for linear elasticity. *SIAM J. Sci. Comput.* 2000; 22(4):1199–1219.
- Klawonn A and Widlund OB. FETI and Neumann-Neumann iterative substructuring methods: connections and new results. *Commun. Pure Appl. Math.* 2001; 54:57–90.
- Klawonn A and Widlund OB. Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients. *SIAM J. Numer. Anal.* 2002; 40(1):159–179.
- Korneev VG. Almost optimal solver for Dirichlet problems on subdomains of decomposition. *Differ. Equations* 2001; 37(7):958–968 (in Russian).
- Korneev VG. Local Dirichlet problems on subdomains of decomposition in hp discretizations, and optimal algorithms for their solution. *Math. Model.* 2002; 14(5):51–74.
- Korneev VG and Jensen S. Preconditioning of the p -version of the finite element method. *Comput. Methods Appl. Mech. Eng.* 1997; 150(1–4):215–238.
- Korneev VG and Jensen S. Domain decomposition preconditioning in the hierarchical p -version of the finite element method. *Appl. Numer. Math.* 1999; 29:479–518.
- Korneev VG and Langer U. *Approximate Solution of Plastic Flow Theory Problems*. Teubner Texte zur Mathematik. B.G. Teubner, Leipzig, 1984.
- Korneev V, Langer U and Xanthos L. On fast domain decomposition solving procedures for hp -discretizations of 3- d elliptic problems. *Comput. Methods Appl. Math.* 2003; 3(4):536–559.
- Korneev VG, Flaherty J, Oden T and Fish J. Additive Schwarz algorithms for solving hp -version finite element systems on triangular meshes. *Appl. Numer. Math.* 2002; 43(3):399–421.
- Langer U and Steinbach O. Boundary element tearing and interconnecting methods. *Computing* 2003; 71(3):205–228.
- Lebedev VI and Agoshkov VI. *Poincaré-Steklov Operators and their Applications in Analysis*. Academy of Sciences USSR: Moscow, 1983 (in Russian).
- Le Tallec P. Domain decomposition methods in computational mechanics. In *Computational Mechanics Advances*, Oden JT (ed.). North Holland: New York, 1994; 121–220.
- Lions PL. On the Schwarz alternating method I. In *First International Symposium on Domain Decomposition Methods for Partial Differential Equations*, Glowinski R, Golub GH, Meurant GA and Périaux J (eds). SIAM: Philadelphia, 1988; 1–42.
- Mandel J. Balancing domain decomposition. *Commun. Numer. Methods Eng.* 1993; 9:233–241.
- Mandel J and Brezzina M. Balancing domain decomposition for problems with large jumps in coefficients. *Math. Comput.* 1996; 65:1387–1401.
- Mandel J and Tezaur R. Convergence of a substructuring method with Lagrange multipliers. *Numer. Math.* 1996; 73:473–487.
- Mandel J and Tezaur R. On the convergence of a dual-primal substructuring method. *Numer. Math.* 2001; 88:543–558.
- Matsokin AM and Nepomnyashchikh SV. A Schwarz alternating method in a subspace. *Sov. Math.* 1983; 29(10):78–84.
- Mikhlin SG. On the Schwarz algorithm. *Dokl. Akad. Nauk S.S.S.R.* 1951; 77(4):569–571 (in Russian).
- Morgenstern D. Begründung des alternierenden Verfahrens durch Orthogonalprojektion. *ZAMM* 1956; 36(7/8):255–256 (in German).
- Nepomnyashchikh SV. Fictitious components and subdomain alternating methods. *Sov. J. Numer. Anal. Math. Modell.* 1990; 5(1):53–68.
- Nepomnyashchikh SV. Mesh theorems on traces, normalizations of function traces and their inversion. *Sov. J. Numer. Anal. Math. Modell.* 1991; 6(3):223–242.
- Nepomnyashchikh SV. Method of splitting into subspaces for solving elliptic boundary value problems in complex form domains. *Sov. J. Numer. Anal. Math. Modell.* 1991; 6(2):151–168.
- Nevaniinna R. Über das alternierenden Verfahren von Schwarz. *Crelle's J. Reine Angew. Math.* 1939; 180:121–128 (in German).
- Oden JT, Patra A and Feng YS. Parallel domain decomposition solver for adaptive hp finite element methods. *SIAM J. Numer. Anal.* 1997; 34:2090–2118.
- Ozrag A. Spectral methods for problems in complex geometries. *J. Comput. Phys.* 1980; 37:70–92.
- Oswald P. On discrete norm estimates related to multilevel preconditioners in the finite element method. In *Proceedings of the International Conference on Constructive Theory of Functions*, Ivanov KG, Petrushev P and Sendov B (eds). Varna 1991, Bulg. Acad. Sci.: Sofia, 1992; 203–241.
- Oswald P. *Multilevel Finite Element Approximation: Theory and Application*. Teubner Skripten zur Numerik, B.G. Teubner: Stuttgart, 1994.
- Pavarino LF. Additive Schwarz methods for the p -version finite element method. *Numer. Math.* 1994; 66(4):493–515.
- Pavarino LF and Widlund OB. Polylogarithmic bound for an iterative substructuring method for spectral elements in three dimensions. *SIAM J. Numer. Anal.* 1996; 37(4):1303–1335.
- Przemieniecki JS. Matrix structural analysis of substructures. *AIAA J.* 1963; 1(1):138–147.
- Quarteroni A and Vail A. *Domain Decomposition Methods for Partial Differential Equations*. Oxford Sciences Publications, 1999.
- Schwarz HA. Über einige Abbildungseigenschaften. *J. Reine Angew. Numer. Math.* 1869; 70:105–120.
- Smith BF and Widlund OB. A domain decomposition algorithm using a hierarchical basis. *SIAM J. Sci. Statist. Comput.* 1990; 11(4):1212–1220.
- Smith B, Björstad P and Gropp W. *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge University Press, 1996.

- Sobolev SL. Schwarz algorithm in the theory of elasticity. *Dokl. Akad. Nauk S.S.S.R.* 1936; 4:235–238 (in Russian).
- Stefanica D. A numerical study of FETI algorithms for mortar finite element methods. *SIAM J. Sci. Statist. Comput.* 2001; 23(4):1135–1160.
- Steinbach O. *Stability Estimates for Hybrid Coupled Domain Decomposition Methods*. Springer-Verlag: Berlin, Heidelberg, New York, 2003.
- Tong CH, Chan TF and Kuo CC. A decomposition preconditioner based on a change to a multilevel nodal basis. *SIAM J. Sci. Statist. Comput.* 1991; 12(6):1486–1495.
- Widlund OB. Iterative substructuring methods: algorithms and theory for elliptic problems in the plane. In *First International Symposium on Domain Decomposition Methods for Partial Differential Equations*, Glowinski R, Golub GH, Meurant GA and Périaux J (eds). SIAM: Orlando, Philadelphia, 1988; 113–128.
- Widlund OB. Preconditioners for spectral and mortar finite element methods. In *Proceedings of the Eighth International Conference on Domain Decomposition Methods*, Beijing, PRC, 15–19 May, 1995; Wiley-Interscience: Strasbourg, 1996.
- Wohlmuth B. A mortar finite element method using dual spaces for Lagrange multipliers. *SIAM J. Numer. Anal.* 2000; 38:989–1014.
- Wohlmuth B. *Discretization Methods and Iterative Solvers Based on Domain Decomposition*. Springer: Heidelberg, 2001.

- Xu J. Iterative methods by space decomposition and subspace correction: a unifying approach. *SIAM Rev.* 1992; 34:581–613.
- Xu J and Zikatanov L. The method of alternating projections and the method of subspace corrections in Hilbert space. *J. Am. Math. Soc.* 2002; 15:573–597.
- Xu J and Zou J. Some nonoverlapping domain decomposition methods. *SIAM Rev.* 1998; 40(4):857–914.
- Zhang X. Multilevel Schwarz methods. *Numer. Math.* 1992; 63(4):521–539.

FURTHER READING

- Ben Belgacem F, Seshaiyer P and Suri M. Optimal convergence rates of hp mortar finite element methods for second order elliptic problems. *RAIRO Math. Modeling Numer. Anal.* 2000; 34:591–608.
- Chan TF and Resasco DC. *A Survey of Preconditioners for Domain Decomposition*. Technical Report, DCS/RR-414, Yale University 1985.
- Domain Decomposition home page. <http://www.ddm.org> [30 September 2003].

Chapter 23

Nonlinear Systems and Bifurcations

Werner C. Rheinboldt

University of Pittsburgh, Pittsburgh, PA, USA

| | |
|-------------------------------------|-----|
| 1 Introduction | 649 |
| 2 General Iterative Processes | 650 |
| 3 Some Classes of Iterative Methods | 657 |
| 4 Parameterized Systems | 661 |
| 5 Bifurcation | 669 |
| References | 673 |

1 INTRODUCTION

Nonlinearities pervade all areas of mechanics, and nonlinear systems of equations arise in connection with numerous mechanical problems. The forms and properties of these systems depend strongly on the type of problem and the inherent sources of nonlinearities. It has long been recognized that, except in special cases, direct methods for solving nonlinear systems are unavailable or infeasible and attention must focus on iterative processes. The choice and effectiveness of an iterative technique depends critically on the available information about the particular equations and their solution set, the aims and accuracy requirements of the computation, and the available computational resources. There are generally no satisfactory guidelines for deciding on a 'best' solution process for a class of nonlinear problems. This is not likely to change because of the ever widening range and complexity of problem types under consideration and the continuing advances in scientific computing in response to the fast pace of hardware and software development.

Encyclopedia of Computational Mechanics, Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes. Volume 1: *Fundamentals*. © 2004 John Wiley & Sons, Ltd. ISBN: 0-470-84699-2.

This chapter gives an overview of some of the basic theoretical and numerical results concerning the solution of systems of nonlinear equations and of several topics related to parameterized systems and their bifurcation behavior. Obviously, the presentation can only touch a small part of this extensive area. We begin with a brief notational summary.

1.1 Notations

In order to avoid technical details, this presentation works with n -dimensional real linear spaces \mathbb{R}^n of column vectors x with components x_1, x_2, \dots, x_n . This is not a restriction, since for the most part the material is basis-independent and hence \mathbb{R}^n may also be regarded as an abstract real linear space. Correspondingly, $A \in L(\mathbb{R}^n, \mathbb{R}^m)$ denotes either an $m \times n$ matrix or a linear operator as context dictates. As usual, $GL(n)$ is the general linear group of invertible $A \in L(\mathbb{R}^n)$ ($= L(\mathbb{R}^n, \mathbb{R}^n)$).

With this notation, the nonlinear mappings central to our discussion have the form

$$F: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^m, \quad F(x) := \begin{pmatrix} f_1(x_1, x_2, \dots, x_n) \\ f_2(x_1, x_2, \dots, x_n) \\ \vdots \\ f_m(x_1, x_2, \dots, x_n) \end{pmatrix}$$
$$\forall x := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in E \quad (1)$$

Here, n and m are given dimensions and the domain E is always assumed to be an open subset of \mathbb{R}^n . Multivariable

calculus is an essential tool for the study of nonlinear systems. A mapping (1) is of class C^r , $r \geq 1$, (or a C^r map, for short) on its (open) domain E if for each component functional f_i , $i = 1, \dots, m$, all partial derivatives up to order r exist and are continuous. The mapping is F -differentiable (Frechet-differentiable) at a point $x \in E$ if there exists a linear operator $A \in L(R^n, R^m)$ such that

$$\lim_{h \rightarrow 0} \frac{1}{\|h\|} \|F(x+h) - F(x) - Ah\| = 0$$

The limit is independent of the norms and the unique operator A is called the F -derivative of F at x and denoted by $DF(x)$. With the natural bases used in (1), the derivative has the matrix representation

$$DF(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} f_1(x) & \dots & \frac{\partial}{\partial x_n} f_1(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_m(x) & \dots & \frac{\partial}{\partial x_n} f_m(x) \end{pmatrix}$$

If the F -derivative of F exists at a point x , then F is continuous at that point. A C^1 mapping on E is F -differentiable at each point of E .

2 GENERAL ITERATIVE PROCESSES

2.1 Existence considerations

Meaningful results about: methods for solving a 'square' system

$$F(x) = 0, \quad F: E \subset R^n \rightarrow R^n, \quad E \text{ open} \quad (2)$$

depend critically on the properties of the function F . Two special cases of (2) are better understood than most others, namely, the n -dimensional linear systems (see e.g. Golub and Van Loan, 1989), and the one-dimensional (i.e. scalar) nonlinear equations (see e.g. Heitzinger, Troch and Valentin, 1985). Here, the emphasis will be on methods for the general nonlinear case and not on results that are specific only to these special cases.

The design and study of an effective numerical method for a given system (2) requires some understanding of its solvability properties. Simple scalar examples already show that there may be any finite or infinite number of solutions or none whatsoever. In fact, it has been proved that for each closed set $S \subset R^1$, there exists a C^∞ map from R^1 into itself, which has S as its zero set. This is a special case of a more general n -dimensional result proved by H. Whitney in 1934. If solutions are known to exist for a system (2),

they may exhibit rapid changes under perturbations of F , which usually impede their computation.

The study of the existence of solutions and their properties is a topic of nonlinear functional analysis and is outside the frame of this chapter. We mention only briefly a few of the principal approaches that apply to the finite-dimensional case:

A conceptually simple, but powerful technique is the transformation of (2) into an extremal problem for some nonlinear functional

$$g: E \subset R^n \rightarrow R^1, \quad E \text{ open} \quad (3)$$

A point $x^* \in E$ is a local minimizer of (3) if $g(x) \geq g(x^*)$ for all x in some open neighborhood of x^* in E and a global minimizer if the inequality holds for all $x \in E$. A critical point of g is any $x^* \in E$ where g has an F -derivative for which $Dg(x^*) = 0$. If g is F -differentiable at a local minimizer x^* in the open set E , then x^* is necessarily a critical point. (This result also holds under a weaker differentiability condition.) A mapping (1) is a gradient (or potential) mapping on E if there exists a functional (3) that is F -differentiable on E and satisfies $F(x) = Dg(x)^T$ for all $x \in E$. A C^1 mapping (1) on an open convex set E is a gradient mapping if and only if $DF(x)$ is symmetric for all $x \in E$. For any gradient mapping, the problem of solving the system (2) can be replaced by that of determining the local minimizers of the functional g , provided, of course, we keep in mind that in this way not all solutions of the nonlinear system may be obtained. This corresponds to the variational approach in the theory of differential equations of importance in many areas of mechanics.

Even if F is not a gradient mapping, the system (2) can be converted into a minimization problem. In fact, this also applies to an overdetermined system

$$F(x) = 0, \quad F: E \subset R^n \rightarrow R^m, \quad n \leq m \quad (4)$$

Let $f: R^m \rightarrow R^1$ be a functional that has $x = 0$ as a unique global minimizer. For instance, we might use $f(x) = x^T A x$ with a symmetric, positive-definite $A \in GL(m)$ or $f(x) = \|x\|$ for some norm on R^m . Then each solution $x^* \in E$ of (4) is a global minimizer of the functional $g(x) := f(F(x))$, $x \in E$, and hence, x^* may be found by minimizing g . But note that a global minimizer $x^* \in E$ of g need not satisfy (4) since this system does not even have to have a solution, and very likely for $n < m$, will not have one. Any global minimizer of g on E is called an f -minimal solution of (4). Various cases of f , such as $f(x) = \|x\|_\infty$ or $f(x) = x^T x$, are of special interest. In the latter case, the functional to be minimized has the form $g(x) := F(x)^T F(x)$, $x \in E$. This defines a nonlinear least-squares problem and, correspondingly, the f -minimal

solutions of $F(x) = 0$ are called *least-squares solutions*. In applications, least-squares problems often arise naturally, for example, in the course of estimating parameters in a functional relationship on the basis of experimental data.

Another class of existence results for (2) is based on arguments derived from the contraction principle and its many generalizations. For this, the system is written in the fixed-point form $F(x) := x - G(x)$ involving some mapping G . As the name indicates, the zeros of F are exactly the fixed points of G and the contraction mapping theorem concerns the existence of such fixed points:

Theorem 1. Let $G: E \subset R^n \rightarrow R^n$ satisfy the contraction condition

$$\|G(x) - G(y)\| \leq \alpha \|x - y\|, \quad x, y \in E, \quad \alpha < 1 \quad (5)$$

Then G has a fixed point in every closed subset $C \subset E$ that is mapped into itself by G , that is, for which $GC \subset C$.

Examples show that the contraction property (5) by itself does not suffice to guarantee the existence of a fixed point; in other words, an additional assumption, such as $GC \subset C$, is indeed needed. Theorem 1 holds on complete metric spaces and there are numerous generalizations and extensions; see, for example, Ortega and Rheinboldt (2000).

The contraction condition plays an important role in many parts of multivariable analysis. It underlies, for instance, the familiar inverse and implicit function theorems that provide local existence results for nonlinear equations. Much deeper are the topological approaches used in nonlinear functional analysis ranging from classical degree theory to modern differential topology and global analysis; see, for example, Berger (1977). This includes, for instance, the Brouwer fixed-point theorem, which guarantees that a continuous mapping $G: E \subset R^n \rightarrow R^n$ on a compact, convex set $C \subset E$ has a fixed point in C if it maps C into itself.

2.2 Process characterization

Suppose that for a (square) nonlinear system (2) the existence of solutions has been established. The problem of computing such solutions includes a range of tasks, such as (i) the determination of sets that are known to contain solutions, (ii) the construction of iterative processes for approximating a specific solution, and (iii) the development of methods for determining all solutions. The tasks (i) and (iii) represent as yet wide open research areas. Most of the current methods for (i) are based on the use of interval computations that have become an active and promising research topic in recent years but have not found very widespread use; see, for example, Kearfott (1996).

General methods for (iii) are available only in the case of minimization problems; see, for example, Floudas and Pardalos (1996). Accordingly, the discussion here will focus on methods for the task (ii).

In general, an iterative process \mathcal{J} for approximating a solution of (2) consists of a 'step algorithm' \mathcal{G} for a single iteration step and a 'control algorithm' \mathcal{C} for controlling the course of the iteration. A state of \mathcal{J} is a triple $\{k, x, M\}$ consisting of a step count k , the current iterate $x \in R^n$, and a memory set M . The specific content of the memory set varies with the process type and the implementation details. In particular, M has to provide all the needed information about the procedure for evaluating F and, if needed, the derivatives of F . But, as the name indicates, M may also retain computed data for later use and redundant data that are too costly to recompute. The input of both \mathcal{G} and \mathcal{C} is assumed to be a given state. The evaluation of $\mathcal{G}\{k, x, M\}$ may result in an error, otherwise the output is a new state $\{k, x, M\}$ consisting of the incremented step count $k := k + 1$, the next iterate x , and an updated memory set M . The output of $\mathcal{C}\{k, x, M\}$ consists of one of the three decisions: 'accept', 'fail', and 'continue', signifying a successful completion of the iteration, failure or suspected divergence of the process, and the need for another step respectively. The process is forced to terminate when the iteration count k reaches a specified maximal value k_{\max} . Thus, altogether, \mathcal{J} is an algorithm of the following generic form:

\mathcal{J} input: starting point x , memory set M ,
maximal count k_{\max} ;
 $k := 0$;
while $k < k_{\max}$
 $\text{decision} := \mathcal{C}\{k, x, M\}$;
 if decision = 'accept' **then return** $\{k, x, M\}$;
 if decision = 'fail' **then return** {'process failure'};
 $\text{evaluate } \{k, x, M\} := \mathcal{G}\{k, x, M\}$;
 if 'error' **then return** {'step failed'};
endwhile;
return {'maximal iteration count reached'}; (6)

The process (6) is stationary if the output of \mathcal{G} does not depend on the current value k of the iteration index, otherwise it is nonstationary. If for a fixed $\ell > 0$ and all $k \geq \ell$, the step algorithm \mathcal{G} depends on precisely ℓ prior iterates for the computation of its output, then \mathcal{J} is said to be an ℓ -step method. Obviously, the required prior iterates have to be saved in the memory set. In the simple case $\ell = 1$, we have a one-step process. For a stationary one-step method, the algorithm \mathcal{G} defines a mapping $G: R^n \rightarrow R^n$ and a step of the method can be written in the familiar

form

$$x^{k+1} = G(x^k), \quad k = 0, 1, \dots \quad (7)$$

Since the purpose of the process (6) is the determination of a solution of (2), it is hoped that \mathcal{J} produces iterates x^k that approximate more and more closely a point $x^* \in \mathbb{R}^n$, which then can be verified to satisfy $F(x^*) = 0$. This requires suitable choices of the starting data and k_{\max} such that the control algorithm C produces no failure decisions and ultimately accepts an iterate x^k that can be guaranteed to satisfy $\|x^k - x^*\| \leq \varepsilon$ for a given tolerance $\varepsilon > 0$.

In this generality, little can be proved only because x^* is unknown and the information available during the process is limited and strongly dependent on the problem class, the computer, and the implementation of \mathcal{J} . In general, a convergence analysis of an iterative process becomes possible only if the control algorithm C is disregarded, that is, if \mathcal{J} is allowed to continue indefinitely. There is an extensive literature on such convergence results for various types of methods. We restrict ourselves to the local convergence of simple stationary one-step processes written in the form (7).

2.3 Convergence factors and orders

A comparison of different iterative processes requires some measure of their efficiency and computational cost. One possible approach is to calculate the overall cost of a process as the product of the average cost of executing one step and the estimated total number of steps required for reaching an acceptable iterate. Obviously, both these quantities depend not only on the process and its implementation but also on the specific problem.

The cost of a step includes the cost of executing both the step algorithm \mathcal{G} and the control algorithm C . Around 1960, A. M. Ostrowski proposed the name 'horner' for one unit of this computational work. There is no generally accepted specification of such a unit, but many authors define a horner as one scalar function call in the step algorithm \mathcal{G} . Then, one evaluation of the n components of the mapping F equals n horners and a computation of the first derivative involves maximally n^2 horners.

The number of steps required for reaching an acceptable iterate depends strongly on the chosen starting data and can vary considerably among the sequences $\{x^k\}$ generated by \mathcal{J} . Accordingly, interest has to center on worst-case measures of asymptotic type. The search for definitions of convergence rates of sequences is probably as old as the convergence theory itself. We summarize here an approach by Ortega and Rheinboldt (2000), which is modeled on the standard quotient-test and root-test for infinite series.

For a sequence $\{x^k\} \subset \mathbb{R}^n$ with limit x^* and any $p \in [1, \infty)$, define the R-factors (root-convergence factors) by

$$R_p(x^k) = \begin{cases} \limsup_{k \rightarrow \infty} \|x^k - x^*\|^{1/k}, & \text{if } p = 1 \\ \limsup_{k \rightarrow \infty} \|x^k - x^*\|^{1/p}, & \text{if } p > 1 \end{cases} \quad (8)$$

and the Q-factors (quotient-convergence factors) by

$$Q_p(x^k) = \begin{cases} \limsup_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^p}, & \text{if } x^k \neq x^* \\ 0, & \text{if } x^k = x^* \\ +\infty, & \text{otherwise} \end{cases} \quad (9)$$

Let $C(\mathcal{J}, x^*)$ denote the set of all possible sequences generated by \mathcal{J} that converge to x^* . Since \mathcal{J} has to be assessed by its worst-case behavior, the R-factors and Q-factors of \mathcal{J} with respect to the limit point x^* are defined by

$$\begin{aligned} R_p(\mathcal{J}, x^*) &= \sup \{R_p(x^k) : \{x^k\} \in C(\mathcal{J}, x^*)\}, \\ Q_p(\mathcal{J}, x^*) &= \sup \{Q_p(x^k) : \{x^k\} \in C(\mathcal{J}, x^*)\}, \end{aligned} \quad \forall p \in [1, \infty)$$

respectively. The R-factors and Q-factors have values in $[0, 1]$ and $[0, \infty]$, respectively. The equivalence of all norms on \mathbb{R}^n implies the norm-independence of the R-factors, but simple examples show that the Q-factors depend on the choice of norm. For $p = 1$, the inequality $R_1(\mathcal{J}, x^*) \leq Q_1(\mathcal{J}, x^*)$ always holds, but for $p > 1$ there is no general relation between the R-factors and Q-factors.

The crucial fact about these factors is that they are step functions of p with at most one step from the minimal to the maximal value. In other words, unless $R_p(x^k) = 0$ or $R_p(x^k) = 1$ for all $p \in [1, \infty)$, there exists a $p_0 \in [1, \infty)$ such that $R_p(x^k) = 0$ for $p \in [1, p_0)$ and $R_p(x^k) = 1$ for $p \in [p_0, \infty)$. This value p_0 is called the R-order of the sequence, and accordingly, the R-order of \mathcal{J} at x^* is defined by

$$O_R(\mathcal{J}, x^*) := \begin{cases} \infty & \text{if } R_p(\mathcal{J}, x^*) = 0 \\ \inf\{p \in [1, \infty) : R_p(\mathcal{J}, x^*) = \infty\}, & \text{otherwise} \end{cases}$$

Analogously, unless $Q_p(x^k) = 0$ or $Q_p(x^k) = \infty$ for all $p \in [1, \infty)$, there exists a $p_0 \in [1, \infty)$ such that $Q_p(x^k) = 0$ for $p \in [1, p_0)$ and $Q_p(x^k) = \infty$ for $p \in [p_0, \infty)$. Hence,

the Q-order of \mathcal{J} at x^* is

$$O_Q(\mathcal{J}, x^*) := \begin{cases} \infty & \text{if } Q_p(\mathcal{J}, x^*) = 0 \\ \inf\{p \in [1, \infty) : Q_p(\mathcal{J}, x^*) = \infty\}, & \text{otherwise} \end{cases} \quad \forall p \in [1, \infty)$$

The R-order and Q-order are both norm-independent and satisfy the relation

$$O_Q(\mathcal{J}, x^*) \leq O_R(\mathcal{J}, x^*)$$

With these definitions, it becomes possible to compare the convergence of two iterative processes \mathcal{J}_1 and \mathcal{J}_2 . In terms of the R-measure, this begins with a comparison of the two R-orders $p_1 = O_R(\mathcal{J}_1, x^*)$, $i = 1, 2$: the process with the larger R-order is R-faster than the other one. In the case of equal R-orders $p = p_1 = p_2$, the process with the smaller R-factor $R_p(\mathcal{J}_i, x^*)$ is R-faster. For the Q-measure, the comparison is analogous, except that the comparison of the Q-factors also depends on the choice of norms. The following special terms are often used to characterize the convergence of a process \mathcal{J} at a limit point x^* :

| | |
|--------------------------------------|---------------------------|
| $0 < R_1(\mathcal{J}, x^*) < 1$ | R-linear convergence |
| $0 < Q_1(\mathcal{J}, x^*) < \infty$ | Q-linear convergence |
| $R_1(\mathcal{J}, x^*) = 0$ | R-superlinear convergence |
| $Q_1(\mathcal{J}, x^*) = 0$ | Q-superlinear convergence |
| $O_R(\mathcal{J}, x^*) = 2$ | R-quadratic convergence |
| $O_Q(\mathcal{J}, x^*) = 2$ | Q-quadratic convergence |

2.4 Local convergence

Consider an iterative process (7) defined by a mapping $G: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$. If the sequence of iterates $\{x^k\} \subset E$ converges to a point $x^* \in E$ where G is continuous, then it follows that $x^* = G(x^*)$ and hence that x^* is a fixed point of G . Here, the standard proof of the contraction mapping Theorem 1 provides a first convergence result:

Theorem 2. Under the conditions of Theorem 1, the iterative sequence (7) started from a given point x^0 in the closed set C converges to the (unique) fixed point $x^* \in C$ and the process \mathcal{J} satisfies $R_1(\mathcal{J}, x^*) \leq Q_1(\mathcal{J}, x^*) \leq \alpha$.

Let G be a contraction on its domain E . If $x^0 \in E$ is such that the closed ball

$$\begin{aligned} \bar{B}(G(x^0), r) &:= \{x \in \mathbb{R}^n : \|x - G(x^0)\| \leq r\}, \\ r &= \frac{1}{1 - \alpha} \|G(x^0) - x^0\| \end{aligned} \quad (10)$$

is contained in E . Then G maps this ball into itself and hence Theorem 2 applies. Obviously, the radius is large for α near 1 and for a large first step $\|G(x^0) - x^0\|$. In either case, the ball may not be fully contained in E and an iterate may fall outside of E , causing the process to become undefined and to stop.

The iteration (7) can be viewed as a discrete dynamical process on the domain E . In that terminology, the orbit of a point $x \in E$ is the sequence $\{x^k\}$ defined by $x^0 = x$, $x^{k+1} = G(x^k)$, $k = 0, 1, \dots$. This orbit may terminate at a point $x^{k_0} \notin E$ with a finite index k_0 . Otherwise, the entire sequence $\{x^k\}_{k=0}^\infty$ is defined and remains in E . This is called an infinite orbit. Of course, the iterates x^k of an infinite orbit need not be distinct points of \mathbb{R}^n ; there may well be indices $k_1 < k_2$ such that $x^{k_1} = x^{k_2}$ in which case the sequence of points x^k , $k = k_1, \dots, k_2 - 1$ repeats itself periodically.

For our purposes, interest centers on infinite orbits for which the iterates converge to some point $x^* \in E$. In the case of continuous G , this suggests the following concepts:

Definition 1. For a continuous mapping $G: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ the 'attraction basin' $A(x^*)$ of a fixed point $x^* \in E$ of G is the set of all points $x \in E$ for which the orbit $\{x^k\}$ is infinite and satisfies $\lim_{k \rightarrow \infty} x^k = x^*$. A fixed point $x^* \in E$ of G is a point of attraction of the iteration if it is in the interior of its attraction basin, that is, if x^* has an open 'attraction neighborhood' $U \subset A(x^*)$.

Obviously, for certain fixed points, the attraction basin may consist only of the point itself. We always have $G(A(x^*)) \subset A(x^*)$; but for a point of attraction x^* , the attraction neighborhood U need not be mapped into itself by G .

For an affine mapping $G(x) = Ax + b$ with $A \in L(\mathbb{R}^n)$ and $b \in \mathbb{R}^n$, the attraction basin is fully characterized. In fact, the iteration (7) converges for any starting point $x^0 \in \mathbb{R}^n$ to the unique fixed point $x^* = Ax^* + b$ in \mathbb{R}^n if and only if A has spectral radius $\rho(A) < 1$; that is, if all eigenvalues of A are less than 1 in modulus. In other words, for $\rho(A) < 1$ there exists exactly one fixed point $x^* \in \mathbb{R}^n$ that has the entire space \mathbb{R}^n as its attraction basin. In the nonaffine case, there is only a much weaker result:

Theorem 3. Let $G: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ have a fixed point x^* in the open set E . If G is F -differentiable at x^* , and the spectral radius of $DG(x^*)$ satisfies $\rho(DG(x^*)) < 1$, then x^* is a point of attraction of (7) and $R_1(\mathcal{J}, x^*) = \rho(DG(x^*))$. Moreover, if $\rho(DG(x^*)) > 0$ then $O_R(\mathcal{J}, x^*) = O_Q(\mathcal{J}, x^*) = 1$.

For a proof and for some historical remarks, refer to Ortega and Rheinboldt (2000). Note that in contrast to the affine case, the existence of x^* has to be assumed,

only local instead of global convergence is guaranteed, and merely the sufficiency but not the necessity of the condition $\rho(DG(x^*)) < 1$ is asserted. Counterexamples show that the assumptions of Theorem 3 cannot be improved readily.

If (7) is viewed as a discrete dynamical process, then Theorem 3 represents a result on the asymptotic behavior of the solutions of perturbed linear difference equations at a stationary point. In fact, (7) may be written in the perturbed linear form

$$x^{k+1} - x^* = DG(x^*)(x^k - x^*) + \Phi(x^k), \quad k = 0, 1, \dots$$

where the perturbation $\Phi(x) = G(x) - G(x^*) + DG(x^*)(x - x^*)$ is small in the sense that $\lim_{\|x-x^*\| \rightarrow 0} \Phi(x)/\|x-x^*\| = 0$. In that form, the result (without the convergence rates) was first given in 1929 by O. Perron. Clearly, this provides a close relation to the large body of results on the asymptotic behavior of the solutions of differential equations at stationary points.

If $\rho(DG(x^*)) = 0$ in Theorem 3, then the convergence is R-superlinear. However, it need not be Q-superlinear. The case $DG(x^*) = 0$ permits a stronger conclusion:

Theorem 4. Suppose that the C^1 mapping $G: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ has a fixed point $x^* \in E$ where $DG(x^*) = 0$ and the second F -derivative $D^2G(x^*)$ exists. Then x^* is a point of attraction of (7) and $O_\rho(J, x^*) \geq O_\rho(J, x^*) \geq 2$. Moreover, if $D^2G(x^*)(h, h) \neq 0$ for all $h \neq 0$ in \mathbb{R}^n , then both orders equal 2.

2.5 Attraction basins

In general, the boundaries between attraction basins for different fixed points have a complicated structure and, in fact, exhibit a fractal nature. As an illustration, consider the simple nonlinear system

$$F(x) := \begin{pmatrix} x_1^2 - 3x_1x_2^2 - 1 \\ -x_2^2 + 3x_1^2x_2 \end{pmatrix} = 0 \quad \forall x \in \mathbb{R}^2 \quad (11)$$

representing the real form of the complex cubic equation $z^3 = 1$, $z \in \mathbb{C}$. In terms of x_1, x_2 , the complex Newton method $z^{k+1} = [2(z^k)^3 + 1]/[3(z^k)^2]$ becomes the iterative process

$$x^{k+1} = \frac{2}{3}x^k + \frac{1}{3[(x_1^k)^2 + (x_2^k)^2]^2} \begin{pmatrix} (x_1^k)^2 - (x_2^k)^2 \\ -2x_1^kx_2^k \end{pmatrix} \quad (12)$$

Theorem 3 ensures that each one of the three zeros $(1, 0)^T$, $(-0.5, \pm 0.5\sqrt{3})^T$ of (11) is a point of attraction of (12).

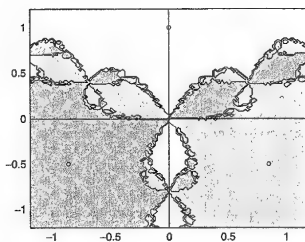


Figure 1. Fractal boundaries between attraction basins.

Figure 1 shows parts of the attraction basins (distinguished by different shadings) of the three zeros (each marked by a small circle). The fractal nature of the boundaries between these basins is clearly visible.

The literature in this area has been growing rapidly. An introduction to fractals is given by Barnsley (1988), and Peitgen and Richter (1986) present interesting graphical examples of attraction basins for various iterative processes. The fractal nature of the boundaries of attraction basins is a property of discrete as well as continuous dynamical systems and has been studied especially in the latter case. For instance, for certain planar differential systems, Nusse and Yorke (1994) showed that there exist basins where every point on the common boundary between two basins and another basin is also on the boundary of a third basin.

These remarks certainly suggest that in an iterative computation of a particular solution x^* of a nonlinear system, it is not unreasonable to expect some very strange convergence behavior unless the process is started sufficiently near x^* . This provides some justification for our emphasis on local convergence results. It also calls for techniques that force the iterates not to wander too far away, and for computable estimates of the radii of balls contained in the attraction basin of a given zero. Some results along this line are addressed in Section 3.

2.6 Acceleration

Basically, Theorems 2 and 3 ensure only the linear convergence of the process (7). Not surprisingly, this has given rise to a large literature on techniques for accelerating the convergence. As Brezinski (1997) shows, much of this is work based on the theory of sequence transformations. Without entering into details, we consider, as an example, the

modification

$$x^{k+1} = vx^k + (1-v)G(x^k), \quad k = 0, 1, \dots \quad (13)$$

of the process (7). Then the following convergence result holds:

Theorem 5. Assume that $G: \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfies

$$\left\{ \begin{aligned} (G(x) - G(y))^T(x - y) &\leq 0, \\ \|G(x) - G(y)\|_2 &\leq \gamma\|x - y\|, \end{aligned} \right\} \quad \forall x, y \in \mathbb{R}^n, \quad \gamma > 0$$

and has a (necessarily unique) fixed point. If $\gamma < 1$, then (13) converges for $v = \gamma^2/(1 + \gamma^2)$ with $R_\gamma(x^k) \leq \sqrt{v}$. If $\gamma \geq 1$, then (13) converges for any $v \in ((\gamma^2 - 1)/(\gamma^2 + 1), 1)$.

In practice, the factor v is often difficult to estimate and is chosen adaptively, that is, (13) is replaced by

$$x^{k+1} = v^k x^k + (1 - v^k)G(x^k), \quad k = 0, 1, \dots$$

Various algorithms for the construction of suitable v^k have been proposed. For example, a generalization of the classical Aitken Δ^2 method leads to Lemaréchal's method

$$v^k = \frac{[G(G(x^k)) - G(x^k)]^T[G(G(x^k)) - 2G(x^k) + x^k]}{[G(G(x^k)) - 2G(x^k) + x^k]^T[G(G(x^k)) - 2G(x^k) + x^k]}$$

In place of (13), two-step accelerations have been considered as well. There are also approaches that lead to transformed sequences with quadratic convergence, but usually they require information that is not easily available in practice; see again Brezinski (1997).

2.7 Condition numbers

An important aspect of any iterative process \mathcal{I} is its sensitivity to perturbations of the mapping F , the problem data, and the implementation. Such perturbations cannot be avoided in computations and examples show that their effects may range from a slowdown of the convergence, to erroneous results, or even to a complete breakdown. As usual, it is desirable to characterize these effects by a single number, the 'condition' of the nonlinear system. We follow an approach by Rheinboldt (1976).

For a mapping $F: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ and any closed set $C \subset E$ set

$$\mu(F, C) := \sup \{ \gamma \in [0, \infty) : \|\tilde{F}(x) - F(y)\| \geq \gamma\|x - y\|, \quad x, y \in C \}$$

$$v(F, C) := \inf \{ \gamma \in [0, \infty) : \|F(x) - F(y)\| \leq \gamma\|x - y\|, \quad x, y \in C \} \quad (14)$$

Then the condition number of F with respect to C is defined by

$$\kappa(F, C) := \begin{cases} \frac{v(F, C)}{\mu(F, C)} & \text{if } 0 < \mu(F, C), v(F, C) < \infty \\ \infty & \text{otherwise} \end{cases} \quad (15)$$

For an affine mapping $F(x) = Ax + b$ with $A \in GL(n)$ and $C = \mathbb{R}^n$, this becomes $\kappa(F, \mathbb{R}^n) = \|A\| \|A^{-1}\|$, that is, the standard condition number $\kappa(A)$ of the matrix A . Suppose that $\kappa(F, C) < \infty$ and that F has a zero $x^* \in C$. Then the definitions (14) imply that F is a homeomorphism from C onto $F(C)$ and hence that $x^* \in C$ is unique. Let $\tilde{F}: \tilde{E} \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ be any perturbation of F with $C \subset \tilde{E}$ and a zero $\tilde{x}^* \in C$. Then, for any given 'reference point' $x^0 \in C$, $x^0 \neq x^*$, the estimate

$$\frac{\|x^* - \tilde{x}^*\|}{\|x^* - x^0\|} \leq \kappa(F, C) \frac{\|\tilde{F} - F\|_C}{\|F(x^0)\|}, \quad \|\tilde{F} - F\|_C = \sup_{x \in C} \|\tilde{F}(x) - F(x)\| \quad (16)$$

holds, which is analogous to a well-known result for linear equations (except that here the existence of the solutions had to be assumed). The estimate (16) affirms that for large $\kappa(F, C)$, the error between the solutions of the original and a perturbed system of equations may become large even if \tilde{F} differs only slightly from F on the set C .

For $\kappa(F, C) < \infty$, it can also be shown that the error of an approximation $y^* \in C$ of the (unique) zero $x^* \in C$ of F satisfies the a posteriori estimate

$$\frac{\|x^* - y^*\|}{\|x^* - x^0\|} \leq \kappa(F, C) \frac{\|F(y^*)\|}{\|F(x^0)\|}, \quad x^0 \in C, \quad x^0 \neq x^* \quad (17)$$

Once again, this corresponds to a familiar linear result. The inequality (17) indicates that if the condition number of F is large, then a small residual norm $\|F(y^*)\|$ does not require the relative error between x^* and y^* to be small as well.

The bounds (14) can be approximated by the norms of derivatives of F . Let $F: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ be continuously differentiable on E and $x \in E$ a point where $DF(x) \in GL(n)$. Then, for sufficiently small $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$|v(F, C) - \|DF(x)\|| \leq \varepsilon, \quad |\mu(F, C) - \|DF(x)^{-1}\|^{-1}| \leq \varepsilon, \quad C := \bar{B}(x, \delta) \subset E$$

and hence

$$\frac{\varepsilon}{\|DF(x)^{-1}\|^{-1} + \varepsilon} \leq \frac{\kappa(F, C) - \kappa(DF(x))}{1 + \kappa(DF(x))} \leq \frac{\varepsilon}{\|DF(x)^{-1}\|^{-1} - \varepsilon} \quad (18)$$

This shows that asymptotically near x the conditions of F and $DF(x)$ are the same and justifies the definition of the matrix condition $\kappa(DF(x)) = \|DF(x)\| \|DF(x)^{-1}\|$ as the pointwise condition number of F at any point $x \in E$ where $DF(x) \in GL(n)$.

This result is of special interest at solutions of $F(x) = 0$. A zero x^* of F is called *simple* if the F -derivative of F at x^* exists and satisfies $DF(x^*) \in GL(n)$. In the scalar case, this reduces to the standard definition of simple roots of differentiable functions. A simple zero $x^* \in E$ of F is locally unique. If F is continuously differentiable in a neighborhood of a simple zero x^* , then we associate with x^* the pointwise condition $\kappa(DF(x^*))$.

At nonsimple zeros of F , the condition results do not apply. Already in the scalar case, it is well known that 'multiple' zeros are difficult to handle and that their character may change quickly with small perturbations of F . Clearly, the condition numbers considered here provide only a partial answer to the wide range of questions relating to the influence of perturbations on a problem and on the corresponding solution processes. The topic still calls for further research.

2.3 Roundoff

As in all numerical computations, the use of finite-precision arithmetic has a potentially profound effect on the solution of nonlinear systems of equations. Roundoff errors are expected in the evaluation of the function (and, where needed, its derivatives), as well as in the execution of the iterative process itself. As all perturbations, these errors may cause a slowdown of the convergence or even a breakdown of the entire process. In addition, it is essential to observe that the roundoff errors in the function evaluation introduce an uncertainty region for the function values.

For any x in the domain E of the function F and a specific floating point evaluation $\hat{F}(x)$ of F , the uncertainty radius at x is defined as $\varepsilon(x) = \|\hat{F}(x) - F(x)\|$. The supremum of the uncertainty radii at the points of a subset $E_0 \subset E$ is the uncertainty radius of F on that set. Ideally, we require that the uncertainty radii are a modest multiple of the unit roundoff of the floating point calculation. This is typically expected of any library function, such as the square-root or the trigonometric functions. But

there are many examples of simple functions F , such as polynomials, where on certain sets a straightforward evaluation of F leads to potentially large uncertainty radii; see, for example, Rheinboldt (1998; Sec. 3.4). Clearly, in such cases there is hardly any hope for an effective iterative determination of the zeros of F . Even if they are not unduly large, the uncertainty radii need to be taken into account in the process. In particular, they play a role in the design of the acceptance and rejection criteria of the control algorithm C .

Let $\{x_d^k\}$ be a sequence of iterates generated by \mathcal{J} (in real arithmetic) that converges to x^* and $\{x_d^k\}$, the corresponding output of a machine implementation of \mathcal{J} using some d -digit arithmetic. Suppose that the control process terminates the sequence $\{x_d^k\}$ with the iterate x_d^k . A satisfactory acceptance test should be expected to guarantee that

$$\lim_{d \rightarrow \infty} x_d^k = x^* \quad (19)$$

As a typical example, consider a frequently used test where the process is terminated at the first index $k^* = k^*(d)$ such that

$$\|x_d^{k^*+1} - x_d^{k^*}\| \leq \varepsilon_d \|x_d^{k^*}\|$$

holds for a given tolerance $\varepsilon_d > 0$. Assume that the process and its implementation are known to satisfy

$$\|x^{k+1} - x^k\| \leq \alpha \|x^k - x^{k-1}\|, \quad \|x_d^k - x^k\| \leq \rho_d, \quad k = 1, 2, \dots$$

with a fixed $\alpha < 1$ and certain roundoff bounds $\rho_d > 0$. Then the convergence of the real sequence implies that

$$\|x_d^{k^*} - x^*\| \leq \frac{1}{1-\alpha} [\varepsilon_d \|x^{k^*}\| + (3 + \varepsilon_d - \alpha) \rho_d] \quad (20)$$

Thus, we need $\lim_{d \rightarrow \infty} \varepsilon_d = 0$ and $\lim_{d \rightarrow \infty} \rho_d = 0$ to prove (19). The choice of the tolerances ε_d is under user control, but the roundoff bounds ρ_d depend strongly on the problem, the iterative process, and the implementation. Here, the uncertainty radii for the function evaluations are coming into play and may cause the roundoff bounds ρ_d to converge to zero much too slowly for any practical purposes. The estimate (20) clearly shows the need for matching the choice of ε_d to that of ρ_d and to the convergence rate of the process (here represented by α). In practice, a mismatch often exhibits itself in an irregular behavior of the computed iterates and their failure to progress satisfactorily toward a solution.

Generally, the design of the acceptance test for a control algorithm C depends strongly on the problem class, the theoretical properties and implementation details of \mathcal{J} , the information available at the time when C is executed, and

the characteristics of the particular computer. It is therefore hardly surprising that most of the known results for such tests, as that of the example, are proved under rather stringent assumptions. Many different types of tests have been proposed, but there are also many examples that show how 'reasonable' tests may fail the criterion (19) for simple problems. This points to the necessity for tailoring a control algorithm, as much as possible, to the specific situation at hand and to consider proving its satisfactory behavior only in that setting. Few results along this line appear to be available in the literature.

3 SOME CLASSES OF ITERATIVE METHODS

3.1 Linearization methods

A major class of iterative methods for solving a (square) nonlinear system (2) is based on the use of linearizations of the mapping F . The idea is to construct at the current iterate $x^k \in E$ an affine approximation

$$L_k: \mathbb{R}^n \rightarrow \mathbb{R}^d, \quad L_k(x) := B_k(x - x^k) - F(x^k), \quad B_k \in GL(n) \quad (21)$$

which agrees with F at x^k , and to use a solution of $L_k(x) = 0$ as the next iterate. Since B_k is assumed to be invertible, the resulting linearization method has the form

$$x^{k+1} = x^k - B_k^{-1} F(x^k), \quad k = 0, 1, \dots \quad (22)$$

In terms of the iterative algorithm (6), this becomes the following step algorithm:

G: **input:** $\{k, x^k, M_k\}$
 evaluate $F(x^k)$;
 construct the matrix B_k ;
 solve $B_k y = F(x^k)$ for y ;
 if solver failed then **return** 'error';
 $x^{k+1} := x^k - y$;
 update the memory set;
return $\{k+1, x^{k+1}, M_{k+1}\}$; (23)

The simplest linearization methods are the (parallel) chord methods where all matrices B_k are identical:

$$x^{k+1} = x^k - B^{-1} F(x^k), \quad k = 0, 1, \dots, \quad B \in GL(n) \quad (24)$$

Typical examples include the Picard iteration with $B = \alpha I$, $\alpha \neq 0$, and the chord Newton method with $B = DF(x^0)$. A special case of the Picard iteration arises if F has the

form $F(x) = Ax - G(x)$ with $A \in GL(n)$ and a nonlinear mapping $G: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$. Then for $B = A$, the chord method (24) becomes $x^{k+1} = A^{-1}G(x^k)$, $k = 0, 1, \dots$, which for $A = I$, that is, for the fixed-point equation $x = G(x)$, is the classical iteration method (7). If F is F -differentiable at x^* , then the iteration function $G(x) := x - B^{-1}F(x)$ of the chord method (24) has at x^* the F -derivative $DG(x^*) := I - B^{-1}DF(x^*)$. Hence, if $\sigma := \rho(I - B^{-1}DF(x^*)) < 1$, then Theorem (3) ensures that x^* is a point of attraction of (24) and $R_1(\mathcal{J}, x^*) = \sigma$.

For differentiable F , the most famous linearization method is Newton's method

$$x^{k+1} = x^k - DF(x^k)^{-1} F(x^k), \quad k = 0, 1, \dots \quad (25)$$

where the affine approximation L_k of (21) is obtained by truncating the Taylor expansion of F at x^k after the linear term.

Another large class of linearization methods with varying B_k are the quasi-Newton methods, also called *Broyden methods* or *update methods*. Here, the matrices are obtained iteratively in the form $B_{k+1} = B_k + \Delta B_k$ where the 'update matrix' ΔB_k has either rank 1 or 2. The theory of these methods began with the formula

$$B_{k+1} = B_k + \frac{F(x_{k+1})(x_{k+1} - x^k)^T}{(x_{k+1} - x^k)^T (x_{k+1} - x^k)}$$

introduced in 1965 by C.G. Broyden. By now the related literature has become very large; for an introduction and references we refer to Kelley (1995) and Dennis and Walker (1981). The methods have been applied to various practical problems such as computational fluid mechanics; see, Engelman, Strang and Bahe (1981).

As discussed in Subsection 2.1, a least-squares solution of an overdetermined problem (4) is a minimizer of the functional $g(x) := F(x)^T F(x)$, $x \in E$. Then a critical point of g is the solution of the system $Dg(x)^T := 2DF(x)^T F(x) = 0$ and hence an application of Newton's method involves the second derivative of F . This can be avoided by approximating g near the k th iterate by the quadratic functional

$$g_k(x) := [F(x^k) + DF(x^k)(x - x^k)]^T [F(x^k) + DF(x^k)(x - x^k)]$$

and then taking a minimizer of g_k as the next iterate x^{k+1} . If $DF(x^k)$ has maximal rank, the global minimizer of g_k is unique and the resulting method becomes

$$x^{k+1} = x^k - [DF(x^k)^T DF(x^k)]^{-1} DF(x^k)^T F(x^k), \quad k = 0, 1, \dots \quad (26)$$

This represents a linearization method for $Dg(x)^T = 0$ and is called the *Gauss-Newton method*. Note that if $m = n$ and $DF(x^k) \in GL(n)$, then (26) reduces to Newton's method for $F(x) = 0$ (but differs from Newton's method applied to $Dg(x)^T = 0$). If $DF(x^k)$ cannot be guaranteed to have maximal rank at all iterates, then a widely accepted approach is to replace (26) by the Levenberg-Marquardt method

$$x^{k+1} = x^k - [\alpha_k I + DF(x^k)^T DF(x^k)]^{-1} DF(x^k)^T F(x^k), \\ k = 0, 1, \dots$$

where the regularization factors $\alpha_k > 0$ are usually determined adaptively.

3.2 Local convergence of Newton's method

Most local convergence results for Newton's method are restricted to simple zeroes of F . For continuously differentiable F , it can be shown that in a neighborhood of a simple zero x^* of F , the iteration function $G(x) := x - DF(x)^{-1}F(x)$ of Newton's method is well defined and that G has the F -derivative $DG(x^*) = 0$ at x^* . Hence, Theorems 3 and 4 provide the following result:

Theorem 6. For a C^1 mapping $F: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$, every simple zero x^* in the (open) set E is a point of attraction of Newton's method (25) with $R_1(J, x^*) = R_1(J, x^*) = 0$. Moreover, there exists $k_0 > 0$ such that

$$\|F(x^{k+1})\| \leq \|F(x^k)\| \quad k \geq k_0 \quad (27)$$

If, in addition, a Lipschitz condition

$$\|DF(x) - DF(y)\| \leq \gamma \|x - y\| \quad \forall x, y \in E, \quad \gamma > 0 \quad (28)$$

holds, then $O_R(J, x^*) \geq O_G(J, x^*) \geq 2$.

The Lipschitz condition (28) can be weakened. In fact, for the convergence at a given simple zero x^* to be at least Q-quadratic, it suffices to assume that

$$\|DF(x) - DF(x^*)\| \leq \gamma \|x - x^*\| \quad (29)$$

in some neighborhood of x^* in E .

An important property of Newton's method is its affine invariance. More specifically, the problem $F(x) = 0$ is invariant under any affine transformation

$$F \longrightarrow \bar{F} := AF, \quad A \in GL(n) \quad (30)$$

and it is easily seen that this invariance is inherited by the Newton iterates. But neither the Lipschitz condition (28) nor its simplified form (29) share this affine invariance. This led P. Deufhard and G. Heindl in 1979 to replace (29) by the affine invariant condition

$$\|DF(x^*)^{-1}[DF(x) - DF(x^*)]\| \leq \gamma \|x - x^*\| \quad (31)$$

The affine invariance of Newton's method is of considerable importance in many situations (see Subsection 3.4).

The following result of Ortega and Rheinboldt (2000) shows that the Newton iteration function is a contraction in some neighborhood of a simple zero, provided the Lipschitz condition (28) holds.

Theorem 7. Assume that the C^1 mapping $F: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfies (28) and has a simple zero $x^* \in E$. Let $\beta = \|DF(x^*)^{-1}\|$ and $\rho > 0$ such that $\eta := \beta\gamma\rho \leq 2/3$. If the closed ball $\bar{B}(x^*, \rho)$ with radius ρ is contained in E , then the Newton iteration function satisfies

$$\|G(x) - x^*\| \leq \frac{1}{2} \frac{\eta}{1 - \eta} \|x - x^*\| \quad \forall x \in B(x^*, \rho)$$

and hence maps $B(x^*, \rho)$ into itself and is a contraction for $\eta < 2/3$.

There are other similar results in the literature. We mention only a convergence theorem for the chord Newton method

$$x^{k+1} = G(x^k), \quad k = 0, 1, \dots, \\ G(x) := x - DF(x^0)^{-1}F(x) \quad (32)$$

which is analogous to the construction of the convergence ball (10) for a contraction:

Theorem 8. Under the assumptions of Theorem 7 for the mapping F , let $x^0 \in E$ be a point where $DF(x^0)$ is invertible. Set $\beta = \|DF(x^0)^{-1}\|$ and assume that $\eta = \beta\gamma\|DF(x^0)^{-1}F(x^0)\| < 1/2$ and that $\bar{B}(x^0, \rho) \subset E$ for $\rho = (1/\beta\gamma)[1 - \sqrt{1 - 2\eta}]$. Then the chord Newton iteration function G of (32) is a contraction on the ball $\bar{B}(x^0, \rho)$ and maps this ball into itself. Hence, the chord Newton method (32) started at x^0 converges to the unique zero x^* of F in $\bar{B}(x^0, \rho)$.

Of course, the convergence follows from Theorem 2. The contraction constant of G on the ball turns out to be $\alpha = 1 - \sqrt{1 - 2\eta}$, which leads to the requirement $\eta < 1/2$. The proof also shows that x^* is a simple zero of F . Theorem 8 is loosely related to the classical theorem for Newton's method proved in 1948 by L.V. Kantorovich.

As in the Kantorovich theorem, it can be shown that x^* is actually a unique zero of F in some larger ball.

3.3 Discretized Newton methods

The execution of step (23) of the linearization method (22) involves (i) the evaluation of the n components of $F(x^k)$, (ii) the evaluation of the n^2 elements of B_k , (iii) the numerical solution of the $n \times n$ linear system, and (iv) the work required in updating the memory set. Evidently, these tasks depend on the problem as well as the method. For instance, in high-dimensional problems, sparse-matrix techniques may be needed in (iii) or direct solvers have to be replaced by iterative methods. For Newton's method, (ii) involves the evaluation of the n^2 first partial derivatives of F . Algebraic expressions for these partial derivatives are not always easily derivable and some automatic differentiation method may be required; see Griewank (2000). Alternatively, the partial derivatives may have to be approximated by appropriate finite differences.

At a point x in the domain of F and for a suitable parameter vector $h \in \mathbb{R}^n$, let $J(x, h) \in L(\mathbb{R}^n)$ be a matrix with elements $J(x, h)_{ij}$ that approximate the partial derivatives $\partial f_i(x)/\partial x_j$, $i, j = 1, \dots, n$. For $J(x, h) \in GL(n)$, the resulting linearization process

$$x^{k+1} = x^k - J(x^k, h^k)^{-1}F(x^k), \quad k = 0, 1, \dots, \quad h^k \in \mathbb{R}^n \quad (33)$$

is called a *discretized Newton method*. A simple example for $J(x, h)$ is

$$J(x, h) \in GL(n), \\ J(x, h)_{ij} = \frac{1}{h_{i,j}} [f_i(x + h_{i,j}e^j) - f_i(x)] \quad (34)$$

where e^1, \dots, e^n are the natural basis vectors of \mathbb{R}^n and $h \in \mathbb{R}^n$ is a vector with small components $h_{i,j} > 0$.

For an analysis of the process (33), we need to know how $J(x, h)$ approximates $DF(x)$. Minimally, the matrix $J(x^k, h^k)$ should be sufficiently close to $DF(x^k)$ at each iterate x^k . The form of J is generally fixed, but the vectors h^k remain free. Evidently, we expect (33) to work only for specific choices of the h^k . Many convergence theories simply assume that $J(x, h)$ is defined for all x in the domain of F and for all h with sufficiently small norm, and that $J(x, h)$ converges to $DF(x)$ when h tends to zero. Certainly, for (34), this is readily proved.

An overview of local convergence results for discretized Newton methods based on various approximation conditions for J is given by Ortega and Rheinboldt (2000). In essence, a simple zero of F is a point of attraction of

(33) if, uniformly at all iterates, the approximation error $\|J(x^k, h^k) - DF(x^k)\|$ is sufficiently small. Moreover, the convergence is superlinear if the approximations tend to zero for $k \rightarrow \infty$. In practice, it is rarely possible to determine a priori estimates for the required approximation properties.

For a given choice of steps $h_{i,j} > 0$, the evaluation of the matrix (34) at a point x requires the function values $f_i(x)$ and $f_i(x + h_{i,j}e^j)$, for $i, j = 1, \dots, n$, which, in the terminology of Subsection 2.2, involves $n + n^2$ homers. Accordingly, not only the approximation properties of J but also its specific form have an effect on the overall behavior of the discretized Newton method. The count in terms of homers is often unrealistic. In fact, in many problems we only have a routine for evaluating the entire vector $F(x) \in \mathbb{R}^n$ at a given point x and the cost of computing a single component value $f_i(x)$ is almost as high as that of computing all of them. In that case, the above example is extraordinarily costly unless we restrict the choice of the steps $h_{i,j}$. For instance, if, say, $h_{i,j} := \bar{h}_j$ for $i = 1, \dots, n$, then in (34) the j th column of $J(x, \bar{h})$ becomes

$$\frac{1}{\bar{h}_j} [F(x + \bar{h}_j e^j) - F(x)] \quad (35)$$

and requires altogether $n + 1$ calls to the routine for F . For sparse Jacobians $DF(x)$, there can be further savings in the evaluation of $J(x, \bar{h})$. For instance, if, for $i = 1, \dots, n$, the component f_i depends at most on x_{i-1} , x_i , and x_{i+1} , then the matrix $J(x, \bar{h})$ with the columns (35) can be generated by four evaluations of F for any dimension $n \geq 3$. An analysis of this approach and relevant algorithms were given by Coleman and Moré (1983).

3.4 Damping strategies

Since the boundaries between the attraction basins of iterative methods are expected to have a fractal nature, the convergence may become erratic unless started sufficiently near the desired solution. For linearization methods, some control of the convergence behavior can be gained by introducing damping factors. We consider here only the representative case of the damped Newton method

$$x^{k+1} = x^k - \lambda_k DF(x^k)^{-1}F(x^k), \quad k = 0, 1, \dots \quad (36)$$

involving certain damping factors $\lambda_k > 0$. Analogous approaches can be used for other linearization methods.

Evidently, for effective control of the iteration, the damping factors should be chosen adaptively. Various strategies have been proposed for this. In the case of a gradient mapping $F(x) = Dg(x)^T$, a popular approach is to construct

λ_k so as to ensure for each step an acceptably large decrease $g(x^k) - g(x^{k+1}) > 0$ in the values of g . Typically, this involves a minimization of g along the line segment $x^k - sp^k$, $0 < s \leq 1$ from x^k in the Newton direction $p^k = DF(x^k)^{-1}F(x^k)$ coupled with a test for the acceptability of the decrease of g . Such a test may require, for instance, that

$$g(x^k - \lambda p^k) \leq (1 - \alpha\lambda)g(x^k) \quad (37)$$

with some constant $\alpha > 0$, say, $\alpha = 1/2$. A possible implementation is here the so-called Armijo rule, where (37) is tested successively with decreasing λ taken from the sequence $\{1, 1/2, 1/4, \dots, \lambda_{\min}\}$. This damping strategy belongs to the topic of unconstrained minimization methods. The literature in this area is extensive; see Rheinboldt (1998) for some references.

In practice, F may not be a gradient mapping or the functional g is not readily computable. Then other functionals have to be constructed that are to decrease at each step. As noted in Subsection 2.4, Newton's method is invariant under affine transformations (30). Moreover, the affine invariant Lipschitz condition (31) suggests the inverse of the Jacobian as a natural choice of the transformation matrix. This leads to the definition of the functional

$$h_k(x) := \frac{1}{2}[DF(x^k)^{-1}F(x)]^T DF(x^k)^{-1}F(x)$$

which is to be decreased in the step from x^k to x^{k+1} . A damping strategy based on this idea was introduced and analyzed by Deufhard (1974). The following algorithm sketches this approach and uses the earlier mentioned Armijo rule to construct a damping factor:

```
input:  $\{x^k, \lambda_{\min}\}$ 
for  $\lambda = 1, \frac{1}{2}, \frac{1}{4}, \dots, \lambda_{\min}$  do
  solve  $DF(x^k)p^k = F(x^k)$ ;
  evaluate  $F(x^k - \lambda p^k)$ ;
  solve  $DF(x^k)q^k = F(x^k - \lambda p^k)$ ;
  if  $\|q^k\|$  is small then return ('convergence detected');
  if  $(q^k)^T q^k \leq (1 - \frac{\alpha}{2})(p^k)^T p^k$  then return ( $\lambda_k = \lambda$ );
endfor;
return ('no  $\lambda_k$  found');
```

The actual implementation of the damping strategy given by Deufhard (1974) is more sophisticated. In particular, the algorithm begins with an a priori estimate of a damping factor that often turns out to be acceptable; moreover, the convergence test takes account of potential pathological situations and the loop uses a recursively computed sequence of λ values. For further details and related convergence aspects, we refer to the original article. The resulting damping strategy is incorporated in the family of Newton codes

NLEQ1, NLEQ2, NLEQ1S available from the Konrad Zuse Zentrum, Berlin.

3.5 Inexact Newton methods

At each step of a linearization method (22), the linear system $B_k(x - x^k) = F(x^k)$ has to be solved. For large dimensions, this may require the use of a secondary linear iterative process such as a successive overrelaxation (SOR) process, alternating direction iterations (ADI), or a Krylov method. We consider here Newton's method as the primary process, in which case the linear system has the form $DF(x^k)s = F(x^k)$. Since any secondary iteration provides only an approximate solution \hat{s} , it cannot be expected that at the next iterate $x^{k+1} = x^k + \hat{s}$ the residual

$$r^k = DF(x^k)(x^{k+1} - x^k) + F(x^k) \quad (39)$$

is zero. The norm $\|F(x^k)\|$ represents a measure of the residual of the primary process. As long as this primary residual is still relatively large, there is little reason for enforcing a very small $\|r^k\|$. Hence, the secondary process should be terminated adaptively on the basis of, for instance, the quotient $\|r^k\|/\|F(x^k)\|$ of the secondary and primary residuals. Combined processes with Newton's method as primary iteration and an adaptive control for the secondary method now carry usually the name *inexact Newton methods* given to them by Dembo, Eisenstat and Steihaug (1982).

It turns out that the convergence of these combined processes can be ensured by requiring the primary residuals $\|F(x^k)\|$ to decrease monotonically. (Recall that Theorem 6 ensures this for Newton's method when k is sufficiently large.) The theory can be based on a convergence result for certain sequences $\{x^k\}$ without taking account of their method of computation. More specifically, for a C^1 mapping $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ and constants $\eta, \lambda \in (0, 1)$, let $\{x^k\} \subset \mathbb{R}^n$ be a sequence such that

$$\|DF(x^k)(x^{k+1} - x^k) + F(x^k)\| \leq \eta \|F(x^k)\| \quad \forall k \geq 0 \quad (40)$$

$$\|F(x^{k+1})\| \leq \lambda \|F(x^k)\| \quad \forall k \geq 0 \quad (41)$$

If a subsequence of $\{x^k\}$ has a limit point x^* where $DF(x^*) \in GL(n)$, then it can be shown that the entire sequence converges to x^* and that $F(x^*) = 0$.

In order to apply this to a combination of Newton's method and some linear iterative method, we have to guarantee the validity of (40) and (41) at each step. This can be accomplished by a damping strategy, that is, a step reduction. At a given point $x \in \mathbb{R}^n$ where $F(x) \neq$

0, and for any step $s \in \mathbb{R}^n$, consider the two control variables

$$r(s) = \frac{\|F(x) + DF(x)s\|}{\|F(x)\|}, \quad \sigma(s) = \frac{\|s\|}{\|F(x)\|}$$

If \hat{s} is a step such that $r(\hat{s}) < 1$, then any reduced step $s = \theta\hat{s}$, $\theta \in (0, 1)$, satisfies

$$r(s) \leq (1 - \theta) + \theta r(\hat{s}) < 1, \quad \sigma(s) = \theta \sigma(\hat{s}) \quad (42)$$

This can be used to show that by means of a step reduction, both (40) and (41) can be satisfied if only (40) already holds for the initial step \hat{s} . The following 'minimum reduction' algorithm given by Eisenstat and Walker (1994) is based on this idea. It involves the choice of suitable parameters $\tau \in (0, 1)$, η_{\max} , $0 < \theta_{\min} < \theta_{\max} < 1$, and j_{\max} , which are assumed to be supplied via the memory sets. A typical acceptance test for methods of this type uses the residual condition $\|F(x^k)\| \leq \text{tol}$ with an appropriate tolerance.

```
G: input:  $\{k, x^k, M_k\}$ 
  apply the secondary process to determine  $s \in \mathbb{R}^n$ 
  such that
   $\|F(x^k) + DF(x^k)s\| \leq \eta \|F(x^k)\|$  for some
   $\eta \in (0, \eta_{\max})$ ;
   $j := 0$ ;
  while  $\{\|F(x^k + s)\| > [1 - \tau(1 - \eta)]\|F(x^k)\|\}$ 
  choose  $\theta \in [\theta_{\min}, \theta_{\max}]$ ;
   $\eta := (1 - \theta) + \theta\eta$ ;  $s := \theta s$ ;  $j := j + 1$ ;
  if  $j > j_{\max}$  then return ('fail');
  endwhile;
   $x^{k+1} = x^k + s$ ;
  return  $\{x^{k+1}, M_{k+1}\}$ 
```

The initial step s satisfies $r(s) \leq \eta < 1$, and since $1 - \eta$ is reduced by a factor $\theta \leq \theta_{\max} < 1$ during each repetition of the while loop, it is expected that the condition of the loop will be achieved for appropriately chosen θ_{\min} and j_{\max} . If the algorithm (43) does not fail and the computed sequence $\{x^k\}$ has a subsequence that converges to a point x^* where $DF(x^*) \in GL(n)$, then it follows from the earlier indicated convergence result that the entire sequence converges to x^* and that x^* is a simple zero of F .

The algorithm (43) requires a secondary iterative method for computing an approximate solution s of the linear system $DF(x^k)s = F(x^k)$ such that $r(s) < 1$. For this, the general minimum residual method (GMRES) has been widely used, although, in principle, any other linear iterative process can be applied. We refer to Kelley (1995) for an introduction to Krylov-type methods and GMRES.

Many variations of inexact Newton methods have been discussed in the literature. For example, Deufhard and Weiser (1998) consider nonlinear elliptic partial differential equations in a finite element setting and combine Newton's method with a multigrid method as a secondary process for the approximation of the Newton corrections. For high-dimensional problems, parallel computations are indicated. For this high-latency, low-bandwidth workstation clusters are increasingly used. A combined process of Newton-Krylov-Schwarz type has been implemented in 1995 at ICASE, Hampton, VA, and used in various aerodynamic applications.

4 PARAMETERIZED SYSTEMS

Nonlinear equations in applications almost always involve several parameters. While some of them can be fixed, for many others, only a possible range is known. Then interest centers on detecting parameter configurations in which the solution behavior exhibits major changes as, for instance, where a mechanical structure starts buckling.

Problems of this type require the incorporation of the changeable parameters into the specification of the equations and hence lead to equations of the form

$$F(y, \lambda) = 0, \quad F: E \subset \mathbb{R}^n := \mathbb{R}^m \times \mathbb{R}^d \rightarrow \mathbb{R}^m, \quad (44)$$

$$n = m + d, \quad d > 0$$

involving a state vector $y \in \mathbb{R}^m$ and a parameter vector $\lambda \in \mathbb{R}^d$. For such systems, it is often convenient to disregard the splitting $\mathbb{R}^n = \mathbb{R}^m \times \mathbb{R}^d$ by combining y and λ into a single vector $x \in \mathbb{R}^n$ and writing (44) in the 'underdetermined' form

$$F(x) = 0, \quad F: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^m, \quad n - m = d > 0 \quad (45)$$

For equations of the form (45) (or (44)), it rarely makes sense to focus only on the computation of a specific solution. Instead, the aim is to analyze the properties of relevant parts of the solution set of the system.

4.1 Homotopies and piecewise linear methods

We begin with a class of techniques, the so-called homotopy methods, in which a parameter is not intrinsic to the problem but is introduced as an aid to the analysis and the computation.

Two continuous mappings $F_i: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$, $i = 1, 2$, are homotopic if there exists a continuous mapping $H: E_H \subset \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^m$ such that $E \times \{0, 1\} \subset E_H$ and

$H(y, 0) = F_0(y)$ and $H(y, 1) = F_1(y)$ for all $y \in E$. A principal question of homotopy theory concerns the preservation of the solvability properties of the mappings $y \in E \mapsto H(y, \lambda)$ when λ changes from 0 to 1. Such preservation information represents a powerful tool for the establishment of existence results and the development of computational methods.

Suppose that a zero y^* of F_1 is to be found. Since, in practice, iterative solvers can be guaranteed to converge to y^* only if they are started sufficiently near that point, techniques are desired that 'globalize' the process. For this, let H be a homotopy H that 'connects' F_1 to a mapping F_0 for which a zero y^0 is already known. If the homotopy preserves the solvability properties, then we expect the solutions $y = y(\lambda)$ of the intermediate systems $H(y, \lambda) = 0$ to form a path from y^0 to y^* . The idea of the homotopy methods is to reach y^* by following this path computationally. Subsections 4.3 and 4.4 address some methods for approximating such solution paths.

Homotopies arise naturally in connection with simplicial approximations of mappings. A basic result of algebraic topology states that a continuous mapping F between certain subsets of finite-dimensional spaces can be approximated arbitrarily closely by a simplicial mapping that is homotopic to F . The details of this result are outside the frame of this presentation. In the past decades, these concepts and results have been transformed into effective computational algorithms for approximating homotopies by simplicial maps on simplicial complexes and their subdivisions. They are now generally called *piecewise linear methods*. The first computational algorithms utilizing simplicial approximations were given in 1964 by C.E. Lemke and J.T. Howson, and addressed the numerical solution of linear complementarity problems. In 1967, H.E. Scarf introduced algorithms for approximating fixed points of continuous mappings that were later shown to belong to this class of methods as well. Since then, the literature on piecewise linear methods has grown rapidly.

Applications of the piecewise linear methods, besides those already mentioned, include the solution of economic equilibrium problems, certain integer programming problems, the computation of stationary points of optimization problems on polytopes, and the approximate solution of continuation problems. The methods also provide a computationally implementable proof of the Brouwer fixed-point theorem mentioned in Subsection 2.1. A survey of the area with extensive references was recently given by Allgower and Georg (2000).

The piecewise linear methods do not require smooth mappings and hence have a theoretically broad range of applicability. In fact, they have also been extended to the computation of fixed points of set-valued mappings. But

usually, these methods are considered to be less efficient when more detailed information about the structure and smoothness of F permits the utilization of other techniques. This appears to be one of the reasons these methods have not found extensive use in computational mechanics.

4.2 Manifolds

Most mechanical applications leading to parameterized equations (44) involve mappings F that are known to belong to some smoothness class C^r , $r \geq 1$. Then typically, the solution set $\mathcal{M} := F^{-1}(0)$ has the structure of a differentiable submanifold of \mathbb{R}^n . In mechanical equilibrium studies, this is often reflected by the use of the term *equilibrium surface*, although a mathematical characterization of the manifold structure of \mathcal{M} is rarely provided.

We summarize briefly some relevant definitions and results about submanifolds of \mathbb{R}^n and refer for details to the standard textbooks. A C^1 mapping (1) on the open set E is an immersion or submersion at $x^0 \in E$ if $DF(x^0) \in L(\mathbb{R}^n, \mathbb{R}^m)$ is a one-to-one mapping or a mapping onto \mathbb{R}^m respectively. The mapping F is an immersion or submersion on a subset $S \subset E$ if it has that property at each point of S . With this, submanifolds of \mathbb{R}^n can be defined as follows:

Definition 2. A subset $M \subset \mathbb{R}^n$ is a d -dimensional C^r submanifold of \mathbb{R}^n , $r \geq 1$, if M is nonempty and for every $x_0 \in M$ there exists an open neighborhood $U \subset \mathbb{R}^n$ of x_0 and a submersion $F: U \rightarrow \mathbb{R}^m$ of class C^r such that $M \cap U = F^{-1}(0) := \{x \in U : F(x) = 0\}$.

The following special case is frequently used:

Theorem 9. Let $F: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$, $n - m = d > 0$, be of class C^r , $r \geq 1$, on an open set E and a submersion on its zero set $M := F^{-1}(0)$. Then M is either empty or a d -dimensional C^r submanifold of \mathbb{R}^n .

For the analysis of submanifolds, local parameterizations are needed.

Definition 3. A local d -dimensional C^r parameterization of a nonempty set $M \subset \mathbb{R}^n$ is a pair (U, φ) consisting of an open set $U \subset \mathbb{R}^d$ and a C^r mapping $\varphi: U \rightarrow \mathbb{R}^n$ such that $\varphi(U)$ is an open subset of M (under the induced topology of \mathbb{R}^n), and φ is an immersion on U that maps U homeomorphically onto $\varphi(U)$. The pair (U, φ) is called a local parameterization near the point x if $x \in M \cap \varphi(U)$.

A nonempty subset $M \subset \mathbb{R}^n$ is a d -dimensional C^r submanifold of \mathbb{R}^n if and only if M has a d -dimensional C^r local parameterization near each of its points.

Let M be a d -dimensional C^r submanifold of \mathbb{R}^n and suppose that at a point $x \in M$ the pair (U, φ) is as

stated in Definition 2. Then the d -dimensional subspace $\ker DF(x) := \{h \in \mathbb{R}^n : DF(x)h = 0\}$ is independent of the specific choice of the local submersion F and depends only on M and the particular point. This linear space is the 'tangent space' of M at x and is denoted by $T_x M$. The subset $TM := \bigcup_{x \in M} \{x\} \times T_x M$ of $\mathbb{R}^n \times \mathbb{R}^n$ is the 'tangent bundle' of M .

Every open subset $E \subset \mathbb{R}^n$ is an n -dimensional C^∞ submanifold of \mathbb{R}^n that has $TE = E \times \mathbb{R}^n$ as its tangent bundle. In particular, the tangent bundle of \mathbb{R}^n itself is $T\mathbb{R}^n = \mathbb{R}^n \times \mathbb{R}^n$. Thus, the tangent bundle of a submanifold M of \mathbb{R}^n appears as a subset of the tangent bundle $T\mathbb{R}^n$ and is itself a submanifold of $T\mathbb{R}^n$ if F is sufficiently smooth. In fact, if M is a d -dimensional C^r submanifold of \mathbb{R}^n with $r \geq 2$, then TM is a $2d$ -dimensional C^{r-1} submanifold of $T\mathbb{R}^n$. Moreover, if $(x, v) \in TM$ and (U, φ) is a local C^r parameterization of M near x , then the pair $(U \times \mathbb{R}^d, (\varphi, D\varphi))$ is a local C^{r-1} parameterization of TM near (x, v) .

The computation of local parameterizations of a manifold M utilizes the following concept:

Definition 4. A d -dimensional linear subspace T of \mathbb{R}^n is a local coordinate space at the point $x \in M$ if

$$T_x M \cap T^\perp = \{0\} \quad (46)$$

If (46) fails to hold, then x is a foldpoint of M with respect to T .

Evidently, at $x \in M$, the tangent space $T = T_x M$ is an obvious choice of a local coordinate space. The canonical inner product of \mathbb{R}^n induces on M a Riemannian structure and it makes sense to introduce the normal spaces

$$N_x M := T_x M^\perp \quad \forall x \in M \quad (47)$$

Then (46) can be written as $N_x M \cap T = \{0\}$.

The computation of a local parameterization on a submanifold is a local process and hence it is no restriction to phrase the following result in terms of the zero set of a single submersion:

Theorem 10. With the assumptions of Theorem 9, let $T \subset \mathbb{R}^n$ be a coordinate subspace of M at $x^c \in M$ and V an $n \times d$ matrix such that the columns form an orthonormal basis of T . Then the C^r mapping

$$H: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad H(x) = \begin{pmatrix} F(x) \\ V^T(x - x^c) \end{pmatrix}, \quad x \in E \quad (48)$$

is a local diffeomorphism from an open neighborhood of $x^c \in M$ onto an open neighborhood U of the origin of \mathbb{R}^d

and the pair (U, φ) defined with

$$\varphi(y) \in \mathbb{R}^d \mapsto M, \quad \varphi(y) := H^{-1} \begin{pmatrix} 0 \\ y \end{pmatrix}, \quad y \in U$$

is a local C^r parameterization of M near x^c .

Thus, the evaluation of $x = \varphi(y)$ for $y \in U$ requires the solution of a nonlinear system of equations. By (46), the derivative $DH(x^c)$ is invertible and experience has shown that a chord Newton method with this derivative works well in practice. The process can be applied in the form

$$x^{k+1} := x^k - DH(x^k)^{-1} \begin{pmatrix} F(x^k) \\ 0 \end{pmatrix}, \quad x^0 = x^c + Vy \quad (49)$$

where the y -dependence occurs only at the initial point. A possible implementation is given by the following algorithm:

```

input { $x^c, y, V, DF(x^c)$ , tolerances}
 $x := x^c + Vy$ ;
compute the LU factorization of  $DH(x^c)$ ;
while {'iterates do not meet tolerances'}
  evaluate  $F(x)$ ;
  solve  $DH(x^c)w = \begin{pmatrix} F(x) \\ 0 \end{pmatrix}$ ;
   $x := x - w$ ;
endwhile;
return  $\{\varphi(y) := x\}$ 

```

In order to meet the condition (46) for the coordinate space T , it is useful to begin with a complementary subspace $S \subset \mathbb{R}^n$ of the tangent space $T_x M$ and to set $T = S^\perp$. If z^1, \dots, z^m is a basis of S , then T is the nullspace of the $m \times n$ matrix Z of rank m with these vectors as its rows. A well-known approach for a nullspace computation is based on the LQ -factorization (with row pivoting) $Z = P^T(L \ 0)Q^T$. Here, P is an $m \times m$ permutation matrix, L an $m \times m$ nonsingular lower-triangular matrix, and $Q = (Q_1 \ Q_2)$ an $n \times n$ orthogonal matrix partitioned into an $n \times m$ matrix Q_1 and an $n \times d$ matrix Q_2 . Then the d columns of Q_2 form the desired orthonormal basis of T . This justifies the following algorithm:

```

input { $Z$ }
compute  $LQ$  factorization of  $Z$  using row pivoting;
for  $j = 1, 2, \dots, d$  do  $u^j := Q_2 e^j$ ;
return  $\{U := (u^1, \dots, u^d)\}$ 

```

Other algorithms for the computation of nullspace-bases of $m \times n$ matrices have been discussed in the literature; see, for example, Rheinboldt (1998) for references.

Instead of choosing the subspace S , we construct the basis matrix Z directly from the derivative $DF(x^*)$. For example, if the coordinate space T is to contain, say, the i th canonical basis vector e^i of \mathbb{R}^n , then Z is formed by zeroing the i th column of $DF(x^*)$ provided, of course, the resulting matrix still has rank m so that (46) holds. Obviously, when the tangential coordinate system is used at x^* , then (51) can be applied directly to $DF(x^*)$ as the matrix Z .

These algorithms represent a small part of the MANPACK package of FORTRAN programs developed by Rheinboldt (1996), which is available on netlib.org.

4.3 Continuation by differentiation

Continuation methods concern problems of the form (45) with $d = 1$, that is, systems

$$F(x) = 0, \quad F: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^m, \quad n = m + 1 \quad (52)$$

defined by a C^r , $r \geq 1$, mapping F on an (open) set E . If F is a submersion on its solution set $\mathcal{M} := F^{-1}(0)$, then \mathcal{M} is a one-dimensional submanifold of \mathbb{R}^n . This manifold may have several connected components, each of which is known to be diffeomorphic either to the unit circle in \mathbb{R}^2 or to some interval on the real line, that is, to a connected subset of \mathbb{R} that is not a point.

Thus, continuation methods can be viewed as methods for approximating connected components of one-dimensional manifolds defined by a nonlinear system (52). There are several different approaches for designing such methods as is reflected by some of the alternate names that have been used including, for instance, embedding methods, homotopy methods, parameter variation methods, and incremental methods. In essence, there are three principal classes of methods, namely, (i) piecewise linear algorithms, (ii) continuation by differentiation, and (iii) methods using local parameterizations.

Piecewise linear continuation algorithms have been developed for computing a simplicial complex of dimension n that encloses the manifold \mathcal{M} . We refer to the survey by Allgower and Georg (1997) for details. Since the simplices of the complex have the dimension n of the ambient space, the methods are best suited for problems in relatively low-dimensional spaces and hence have found only limited application in computational mechanics.

The methods in class (iii) include, in particular, the continuation methods used extensively in computational mechanics and will be the topic of Subsection 4.4.

For the methods in class (ii), consider first the case in which the solution set of (52) can be parameterized in terms

of λ . In other words, assume that there is a C^1 mapping $\eta: \mathcal{J} \rightarrow E$ on an interval \mathcal{J} such that

$$F(\eta(\lambda), \lambda) = 0 \quad \forall \lambda \in \mathcal{J} \quad (53)$$

Let $y^0 = \eta(\lambda^0)$, $\lambda^0 \in \mathcal{J}$ be a given point. Then $y = \eta(\lambda)$ is on \mathcal{J} a solution of the initial value problem

$$D_y F(y, \lambda)y' + D_\lambda F(y, \lambda) = 0, \quad y(\lambda^0) = y^0 \quad (54)$$

Conversely, for any solution $y = \eta(\lambda)$ of (54) on an interval \mathcal{J} such that $\lambda^0 \in \mathcal{J}$ and $F(y^0, \lambda^0) = 0$, the integral mean value theorem implies (53).

In a lengthy series of papers during a decade starting about 1952, D. Davidenko showed that a variety of nonlinear problems can be solved by embedding them in a suitable parameterized family of problems and then integrating the corresponding ODE (54) numerically. He applied these 'embedding methods' not only to nonlinear equations but also to integral equations, matrix inversions, determinant evaluations, and matrix eigenvalue problems; see Ortega and Rheinboldt (2000) for some references. In view of this work, the ODE (54) has occasionally been called the 'Davidenko equation'.

If $D_y F(y, \lambda)$ is nonsingular for all $(y, \lambda) \in E$, then classical ODE theory ensures the existence of solutions of (54) for every $(y^0, \lambda^0) \in E$. But if $D_y F(y, \lambda)$ is singular at certain points of the domain, then (54) is an implicit ODE near such points and the standard theory no longer applies. This was never addressed by Davidenko. But the difficulty can be circumvented by dropping the assumption that the solution set can be parameterized in terms of λ . More specifically, the following result holds:

Theorem 11. Suppose that the C^1 mapping $F: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^{n-1}$ is a submersion on E . Then for each $x \in E$ there exists a unique vector $u_x \in \mathbb{R}^n$ such that

$$DF(x)u_x = 0, \quad \|u_x\|_2 = 1, \quad \det \begin{pmatrix} DF(x) \\ u_x \end{pmatrix} > 0 \quad (55)$$

Moreover, the mapping $\Psi: E \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ defined by $x \in E \mapsto \Psi(x) := u_x$ is locally Lipschitz continuous on E .

Note that F is required to be a submersion on its domain E and not merely on its zero set. This is not a severe restriction and can be ensured by shrinking E if needed. By the local Lipschitz continuity of the mapping Ψ , ODE theory guarantees that for every $x^0 \in E$ the autonomous initial value problem

$$\frac{dx}{dt} = \Psi(x), \quad x(0) = x^0 \in E \quad (56)$$

has a unique solution $x: \mathcal{J} \rightarrow E$ on an open interval \mathcal{J} with $0 \in \mathcal{J}$. This solution can be extended to an interval \mathcal{J} that is maximal with respect to set inclusion. Then, at a finite boundary value $s \in \partial\mathcal{J}$ we have either $x(s) \rightarrow \partial E$ or $\|x(s)\|_2 \rightarrow \infty$ as $s \rightarrow s$, $s \in \mathcal{J}$. A solution $x = x(\tau)$ of (56) satisfies $DF(x(\tau))x'(\tau) = DF(x(\tau))\Psi(x(\tau)) = 0$ whence $F(x(\tau)) = F(x^0)$ for $\tau \in \mathcal{J}$ and therefore $x(\tau) \in F^{-1}(0)$ provided x^0 was chosen such that $F(x^0) = 0$.

These results show that standard ODE solvers can be applied to (56) for computing connected segments of $F^{-1}(0)$. For this, a routine is needed for evaluating $\Psi(x)$ at a given point $x \in E$. A typical approach is to calculate a nonzero vector $u \in \ker DF(x)$, which is then normalized to Euclidean length 1 and oriented by multiplication with a suitable $\sigma = \pm 1$. The direct implementation of the determinant condition in (55) can be avoided by choosing σ such that $\sigma u^T \Psi(x) \geq 0$ where $\Psi(x)$ is the computed vector at some 'earlier' point \tilde{x} .

During the numerical integration of (56) starting from $x^0 \in F^{-1}(0)$, the condition $F(x) = 0$ is not explicitly enforced and the computed points may drift away from $F^{-1}(0)$. Thus, the ODE approach requires further corrections if the aim is to generate a good approximation of the manifold. However, for the homotopy methods introduced in Subsection 4.1, the drift is entirely acceptable. In fact, in the setting of homotopies, interest centers on computing a solution y^* of the terminal system $F_1(y) = 0$ and not on approximating a solution path in the zero set of H . Assume that Theorem 11 applies to H . If for given $y^0 \in F_0^{-1}(0)$ the numerical solution of the initial value problem (56) for H reaches the plane $L := \{(y, \lambda): \lambda = 1\} \subset \mathbb{R}^m \times \mathbb{R}$ at a point $(\tilde{y}, 1) \in L$, then \tilde{y} is expected to be near y^* . Therefore, a standard iterative process for solving $F_1(y) = 0$ should converge to y^* if started from \tilde{y} .

The condition of reaching L can be deduced from the fact that a maximally extended solution of the initial value problem must leave any compact subset of the domain E_H of H . Let F_0 be chosen such that the solution y^0 of $F_0(y) = 0$ is unique. In addition, assume that the ODE solution starting at y^0 remains bounded and is contained, say, in a cylinder $C := \{(y, \lambda): \|y - y^0\| \leq \delta, 0 \leq \lambda \leq 1\}$. Then, for $C \subset E_H$, it follows that the path has to reach the end $C \cap L$ of the cylinder. For a large enough domain E_H , the boundedness of the path is related to the boundedness of the zero set of H . Thus, a central question in the homotopy approach sketched in Subsection 4.1 is the selection of a homotopy that satisfies Theorem 11.

It was shown by Chow, Mallet-Paret and Yorke (1978) that there are homotopy methods for finding zeros of a nonlinear mapping, which are constructive with probability 1. The theory is based on a parameterized Sard theorem. For this, a family $\tilde{H}: \mathbb{R}^p \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}^n$ of homotopies is

considered that depends on a parameter $\mu \in \mathbb{R}^p$. (Typically, the additional parameter is the initial point.) Let \tilde{H} be sufficiently smooth and a submersion on its zero set. Then the theorem states that for almost all μ^0 , the homotopy $H_{\mu^0}(y, \lambda) := \tilde{H}(\mu^0, y, \lambda)$ is a submersion on its zero set. As indicated above, it can then be concluded that for almost all μ , and hence with probability 1, the solution path of the homotopy H_μ is expected to reach the plane L .

This approach has been the topic of a growing literature on what are now called *probability-one homotopy methods* for solving nonlinear systems. Algorithms for many applications have been proposed, including, especially, for the solution of specific classes of equations, such as polynomial systems and unconstrained minimization problems. A widely used implementation is the software package HOMPACK; see, Watson *et al.* (1997) for the latest FORTRAN 90 version, and further references.

4.4 Local parameterization continuation

Consider an equation of the form (52) defined by a C^r mapping ($r \geq 1$) that is a submersion on its zero set $\mathcal{M} := F^{-1}(0)$. We discuss now continuation methods that utilize local parameterizations of the one-dimensional submanifold \mathcal{M} of \mathbb{R}^n to produce a sequence of points x^k , $k = 0, 1, \dots$, on or near \mathcal{M} starting with a given $x^0 \in \mathcal{M}$. Typically, for the step from x^k to x^{k+1} , a local parameterization of \mathcal{M} near x^k is constructed and a predicted point is determined from which an iterative process, such as (50), converges to an acceptable next point $x^{k+1} \in \mathcal{M}$. A local parameterization can be retained over several steps. Of course, this requires a decision algorithm, which may be based, for example, on the rate of convergence of the iteration at earlier points.

The literature on the design, implementation, and application of these continuation methods is huge and cannot be covered here. A survey from a numerical analysis viewpoint with references until about 1997 was given by Allgower and Georg (1997). Equally extensive, and largely independent, is the literature on continuation methods in engineering. For this, see in particular Chapter 4, Volume 2 by E. Riks in this encyclopedia, where path-following methods and load control for engineering applications are treated in detail.

In line with our overall presentation, we address here only some general mathematical ideas and approaches underlying this class of continuation methods.

For the construction of a local parameterization at x^k , we require a nonzero vector $t^k \in \mathbb{R}^n$ such that (46) holds for $T := \text{span } t^k$. This requires that $t^k \notin \text{rg } DF(x^k)^T$, that is,

that t^k is not orthogonal to the tangent space $T_{x^k}\mathcal{M}$. Two frequent choices are

- (a) $t^k := \pm u^k$, $DF(x^k)u^k = 0$, $\|u^k\| = 1$,
tangent vector
(b) $t^k := \pm e^i$, $e^i \notin N_{x^k}\mathcal{M}$,
basis vector of \mathbb{R}^n

Processes involving (a) are often called *pseudo-arc-length* continuation methods. For the computation of a tangent vector, the algorithm (51) can be used with $Z = DF(x^k)$. Alternatively, with a vector $w \in \mathbb{R}^n$ such that $DF(x^k)w \neq 0$, an (unnormalized) tangent vector \tilde{u} is obtained by solving the augmented system

$$\begin{pmatrix} DF(x^k) \\ w^T \end{pmatrix} \tilde{u} = e^n \quad (58)$$

The vector w should be reasonably parallel to the nullspace of $DF(x^k)$. Often, a canonical basis vector e^i of \mathbb{R}^n is chosen such that the matrix obtained by deleting the i th column of $DF(x^k)$ is nonsingular.

For any choice of the basis vector t^k of the local parameterization, the orientation has to be determined appropriately. A simple and effective approach is to orient t^k for $k > 0$ by a comparison with the direction of the vector t^{k-1} at the previous point x^{k-1} . This presumes that the direction of t^0 is given. In the case of (57a)) it is also possible to apply the orientation definition of (55), which, because of

$$\det \begin{pmatrix} DF(x^k) \\ w^T \end{pmatrix} = [(u^k)^T w] \det \begin{pmatrix} DF(x^k) \\ (u^k)^T \end{pmatrix}$$

can be replaced by a condition on the determinant of the matrix of the system (58).

Once t^k has been chosen, the local parameterization algorithm (50) requires the solution of the augmented nonlinear system

$$\begin{pmatrix} F(x) \\ (t^k)^T(x - x^k) - y \end{pmatrix} = 0 \quad (59)$$

for a given local coordinate $y \in \mathbb{R}^n$. In principle, any one of the iterative solvers for $n \times n$ nonlinear systems can be applied. Besides the chord Newton method utilized in (50), most common are Newton's method, discretized Newton methods, and the Gauss-Newton method. In each case, some form of damping may also be introduced. The selection of the iterative process is of critical importance since it constitutes a major part of the computational cost, especially for large sparse problems. In that case, combined processes for solving the augmented system are often applied, such as inexact Newton methods with a linear solver of Krylov type or some other fast algorithm appropriate for

the problem. Certain linear solvers, such as direct factorization methods, allow for an inexpensive computation of the determinant of the matrix of (58), which, as noted, can be used for orienting the tangent vector. This is not the case for other solvers, such as the Krylov-type methods.

In (50), the iteration starts from a 'predicted point' $x^k + ht^k$. In many continuation codes, other linear predictions $x^k + h_k v^k$ with suitable vectors v^k are used. In particular, v^k is often taken to be the (normalized and oriented) tangent vector at x^k , even if t^k is not the tangent vector. This compares with the Euler method for approximating solutions of ODEs, and hence has been called the *Euler predictor*. For the selection of the step $h_k > 0$ in these linear predictions, a number of algorithms have been proposed. Some of them are modeled on techniques used in ODE solvers, but there are also other approaches. For instance, step selections have been based on information collected in connection with the use of some sufficient decrease criteria in damped Newton methods (see Subsection 3.4). Other algorithms involve the estimation of curvature properties of \mathcal{M} near x^k to gain information about the prediction errors; see, for example, Burkardt and Rheinboldt (1983). Besides linear predictors, extrapolatory predictor algorithms have also been considered. Of course, this requires suitable startup techniques for use at the points where too few earlier points are known.

By their definition, all these continuation methods are intended for the approximation of a connected component of the solution manifold \mathcal{M} . But \mathcal{M} may well have several components, and near certain singularities two components can be close to each other. In such cases, the continuation process may jump unnoticeably from one to the other component. Frequently (but not always), two such components may have opposite orientations, and then the specific algorithm for orienting the basis vector t^k of the local parameterization becomes critical. For instance, if the determinant condition of (55) is used, then the jump between the components shows up as a reversal of the direction of the computed path. This is not the case, for example, if the vector t^k is oriented by a comparison with the direction of t^{k-1} . Such different behavior of two orientation algorithms can be an advantageous tool for the detection of certain types of singular points of parameterized mappings (see Section 5).

4.5 Approximation of higher-dimensional manifolds

Suppose that the mapping F satisfies the conditions of Theorem 9 with $d \geq 2$. Then, continuation methods can be applied for the computation of different paths on the d -dimensional manifold $\mathcal{M} = F^{-1}(0)$. But it is not easy to develop a good picture of a multidimensional manifold

solely from information along some paths on it. In recent years, this has led to the development of methods for a direct approximation of implicitly defined manifolds of dimension exceeding 1. For this, a natural approach is the computation of a simplicial approximation of specified subsets $\mathcal{M}_0 \subset \mathcal{M}$, that is, the construction of a simplicial complex of dimension d in \mathbb{R}^n with vertices on \mathcal{M} such that the points of the carrier approximate \mathcal{M}_0 .

The computation of such triangulations is also required in computer aided geometric design CAGD and related applications. But there the task differs considerably from that encountered here. In fact, in CAGD, the manifolds are typically defined as the image sets of explicitly specified parameterizations. On the basis of such parameterizations, various 'triangulation' methods have been proposed in the literature analogous to those for triangulating domains in linear spaces, for example, in finite element methods.

For manifolds that are defined implicitly as the solution set of nonlinear equations, triangulations have been developed only fairly recently. The earliest work (see Allgower and Georg, 1990) involves the use of piecewise linear methods for the construction of an n -dimensional simplicial complex in the ambient space \mathbb{R}^n that encloses the d -dimensional manifold. The barycenters of appropriate faces of the enclosing simplices are then chosen to define a piecewise linear approximation of the manifold itself. But, in general, the resulting vertices do not lie on the manifold, and, as with other piecewise linear methods, the approach is limited to problems in relatively low-dimensional ambient spaces.

A method for the direct computation of a d -dimensional simplicial complex that approximates an implicitly defined manifold \mathcal{M} in a neighborhood of a point was first given by Rheinboldt (1988). It is based on algorithms that were later incorporated in the MANPACK package of Rheinboldt (1996). By means of smoothly varying local parameterizations (moving frames), standardized patches of triangulations of the tangent spaces of \mathcal{M} are projected onto the manifold. The method is applicable to manifolds of dimension $d \geq 2$ but was used mainly for $d = 2, 3$ have been developed for computing d -dimensional simplicial complexes that approximate specified domains of the manifold; see Rheinboldt (1998) for references.

A different method was introduced in 1995 by R. Melville and S. Mackey. It does not involve the construction of a simplicial complex on an implicitly defined two-dimensional manifold; instead the manifold is tessellated by a complex of nonoverlapping cells with piecewise curved boundaries that are constructed by tracing fish-scale patterns of one-dimensional paths on the manifold. The method

appears to be intrinsically restricted to two-dimensional manifolds. Recently, Henderson (2002) developed another approach, which represents a manifold of dimension $d \geq 2$ as a set of overlapping d -dimensional balls and expresses the boundary of the union of the balls in terms of a set of finite, convex polyhedra.

4.6 Sensitivity

For parameter-dependent problems of the form (44), there is not only interest in computing parts of the solution set but also in determining the sensitivity of the solutions to changes of the parameters. Generally, in the literature, the sensitivity of (44) is defined only near a solution $(y^0, \lambda^0) \in F^{-1}(0)$ where the state y depends smoothly on the parameter vector λ . More specifically, assume that $D_y F(y^0, \lambda^0) \in GL(n)$, and therefore, that the implicit function theorem applies. Then there exists a C^1 mapping $\eta: U \rightarrow E$ on a neighborhood $U \subset \mathbb{R}^d$ of λ^0 such that each point $(y, \lambda) \in F^{-1}(0)$ in a certain neighborhood of (y^0, λ^0) is uniquely specified by $(\eta(\lambda), \lambda)$ for some $\lambda \in U$. With this, the sensitivity of F at (y^0, λ^0) is defined as the derivative $D\eta(\lambda^0)$ and hence is the unique solution of the linear system

$$D_y F(y^0, \lambda^0) D\eta(\lambda^0) = -D_\lambda F(y^0, \lambda^0) \quad (60)$$

This corresponds, of course, to the equation (54) in the setting of Davidenko's embedding methods. If, in practice, the partial derivatives $D_y F$ and $D_\lambda F$ are not accessible, finite difference approximations of these derivatives can be introduced instead. But this calls for estimates of the influence of the approximation errors on the desired solution of (60). Alternatively, in order to determine the sensitivity $D\eta(\lambda^0)\mu$ in the direction of a parameter vector μ , approximations of $D\eta(\lambda^0)\mu$ can be computed by numerical differentiation based on values of $\eta(\lambda^0 + \tau\mu)$ for several values of the scalar τ near $\tau = 0$. Evidently, this sensitivity definition does not reflect any of the underlying geometric aspects of the problem and the indicated approximations utilize little information about the prior computation of (y^0, λ^0) . Accordingly, a sensitivity analysis is often considered to be a 'postprocessing' technique that is applied independently of the way the solution (y^0, λ^0) was found.

For the case in which the solution set $\mathcal{M} := F^{-1}(0)$ is a d -dimensional submanifold of \mathbb{R}^n , a geometric interpretation of the sensitivity concept was introduced by Rheinboldt (1993). In particular, it was shown that a sensitivity analysis can be incorporated effectively in the solution process without an undue increase of the computational cost.

Since the sensitivity concept is local in nature, it is once again no restriction to consider only the zero set of a single

For the characterization of the solution behavior near a bifurcation point, a more selective choice of unfolding is needed. In particular, it is desirable that any perturbation of F of the form $\tilde{F}(\cdot, \cdot) + \epsilon Q(\cdot, \cdot)$, is equivalent, in a certain sense, to $\tilde{F}(\cdot, \cdot, \mu(\epsilon))$ for sufficiently small ϵ . With a specified equivalence definition, unfoldings of this type are called *universal unfoldings*. Most, but not all, bifurcation problems possess such universal unfoldings. A universal unfolding of the pitchfork example (67) is the mapping $(\lambda, \mu_1, \mu_2) \mapsto \lambda^2 - \lambda y - y^2 - \mu_1 - \mu_2 y^2$, where, for given $\mu \in \mathbb{R}^2$ the corresponding solutions in the (y, λ) -plane consist of paths that represent generic perturbations of

the original pitchfork. For details of the theory of universal unfoldings, we refer to Golubitsky and Schaeffer (1985).

In mechanical problems, it is usually difficult to relate specific physical parameters of the problem to the augmenting parameters of a universal unfolding. Accordingly, there is a tendency to work with general unfoldings that are defined by intrinsic physical parameters. For computational purposes, these unfoldings usually suffice.

5.2 Scalar parameter problems

Let (44) be a parameterized nonlinear system with a scalar parameter ($d = 1$) defined by a C^r mapping F with sufficiently large $r \geq 1$. Suppose that there exists an open, connected subset $M_0 \subset F^{-1}(0)$ of the zero set of F with the property that F is a submersion on M_0 . Hence, M_0 is a one-dimensional manifold and $x^* := (y^*, \lambda^*) \in M_0$ is a foldpoint of M_0 (with respect to the parameter) if $D_y F(y^*, \lambda^*)$ is singular, or equivalently, if $(e^*)^T u = 0$ for $u \in T_{x^*} M_0$.

A local parameterization of M_0 consists of an open interval J of \mathbb{R} and a homeomorphism φ from J onto $x(J) := \{(y(s), \lambda(s)) : s \in J\} \subset M_0$. For any $x := x(s)$, $s \in J$, the tangent space $T_x M_0$ of M_0 is spanned by $x'(s)$. Thus, a foldpoint $x^* := x(s^*)$, $s^* \in J$, is a zero of the scalar function

$$\eta: J \rightarrow \mathbb{R}, \quad \eta(s) := (e^*)^T x'(s) \quad (68)$$

A simple zero s^* of η (in the sense that $\eta'(s^*) \neq 0$) is called a *simple foldpoint*. In many applications, such simple foldpoints are of special interest. For example, in certain problems of structural mechanics, λ represents a load intensity and such points may signify the onset of buckling. Numerous methods have been proposed for the computation of simple foldpoints of mappings involving a scalar parameter, and we summarize here only the principal ideas underlying some of them. A numerical comparison of several different methods was given by Melhem and Rheinboldt (1982).

A first class of methods assumes to be near a simple foldpoint $x^* \in M$, and involves the construction of an augmented system of equations that is known to have the desired point as a simple zero. A possible approach is to use a system of the form

$$\begin{pmatrix} F(x) \\ g(x) \end{pmatrix} = 0 \quad (69)$$

with a suitable function $g: U \subset E \rightarrow \mathbb{R}$ that is defined on some neighborhood U of x^* . Possible choices of g include $g(x) := \det D_y F(x)$ and $g(x) := v(x)$ where $v(x)$

denotes the smallest eigenvalue (in modulus) of $D_y F(x)$. Computationally more efficient is the definition

$$g(x) = (e^*)^T A(x)^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad A(x) := \begin{pmatrix} DF(x) \\ (e^*)^T \end{pmatrix} \quad (70)$$

where the index i , $1 \leq i \leq n$ is chosen such that $A(x)$ is invertible. Then a modified Newton method for solving (69) involves the following step algorithm:

G: input: $\{k, x^k, M_k\}$
determine i , $1 \leq i \leq n$ such that $A(x^k)$ is invertible;
solve $A(x^k)u^1 = (F(x^k), 0)^T$ and $A(x^k)u^2 = (0, 1)^T$;
 $x^{k+1} := x^k - u^1 + [(e^*)^T(x^k - u^1)/(e^*)^T u^2] u^2$;
update the memory set;
return $\{k+1, x^{k+1}, M_{k+1}\}$; (71)

Initially, a suitable index i has to be supplied and usually $i = n$ is used. At subsequent steps, i is selected such that $|(e^*)^T u^1|$ is maximal for the last computed u^1 .

Instead of an n -dimensional system (69), a larger system of the form

$$\begin{pmatrix} F(x) \\ D_y F(x)w \\ b^T w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad x \in E, \quad w \in \mathbb{R}^m \quad (72)$$

can also be used where $b \in \mathbb{R}^m$ is such that the matrix in the last two rows is invertible. Usually, some canonical basis vector of \mathbb{R}^m is chosen here. If (72) is to be solved by a form of Newton's method, then second derivative terms $D_{yy} F$ and $D_{y\lambda} F$ are needed. This can be avoided by replacing these terms by finite difference approximations. For example, Moore and Spence (1980) suggested the approximation

$$D_{yy} F(x)(h^1, h^2) \approx \frac{1}{\delta} [D_y F(y + \delta h^1, \lambda) h^2 - D_y F(y, \lambda) h^2]$$

with some small $\delta \neq 0$ and showed that the $(n+m)$ -dimensional linear systems arising at each iteration step can be partitioned such that only four linear systems of dimension m with the same matrix have to be solved.

A second class of methods for the computation of simple foldpoints assumes that a specific continuation process is used to approximate a connected, one-dimensional solution manifold M_0 . Since simple foldpoints are simple zeros of the scalar mapping (68), a foldpoint is expected to have been passed during the continuation process when

$$\text{sign}(e^*)^T u^k \neq \text{sign}(e^*)^T u^{k+1}, \quad T_{x^*} M_0 = \text{span}(u^j), \\ j = k, k+1 \quad (73)$$

at two succeeding points x^k, x^{k+1} computed in the process. Hence, in terms of the local parameterization, we want to compute s^* as a zero of the scalar function (68). Here, any one of the standard algorithms for solving scalar equations can be applied. A frequently used algorithm is the well-known Brent method, which combines root bracketing, bisection, and inverse quadratic interpolation.

At a simple foldpoint, the condition $(e^*)^T x'(s^*) = 0$ implies that the scalar function $s \in J \mapsto \zeta(s) := (e^*)^T x(s)$ has an extremum at $s = s^*$. Hence, s^* can also be determined by an algorithm for maximizing or minimizing ζ . For this, various interpolatory methods turn out to be very effective; see again Melhem and Rheinboldt (1982) for details and references.

So far, F was assumed to be a submersion at all points of an open, connected subset of the zero set $F^{-1}(0)$. We now allow this submersion condition to fail at certain points. More specifically, consider a path in $F^{-1}(0)$ defined by a C^r mapping, $r \geq 1$,

$$\pi: J \subset \mathbb{R} \rightarrow \mathbb{R}^n, \quad F(\pi(s)) = 0, \quad \pi'(s) \neq 0 \quad \forall s \in J$$

on an open interval J . Let $x^* := \pi(s^*)$, $s^* \in J$, be a bifurcation point of F where $\text{rank } DF(x^*) = m-1$. Moreover, assume that x^* is an isolated singularity and, for simplicity, assume that $\text{rank } DF(\pi(s)) = m$ for $s \in J$, $s \neq s^*$. Then the orientation function

$$\delta: J \rightarrow \mathbb{R}, \quad \delta(s) := \det \begin{pmatrix} DF(\pi(s)) \\ \pi(s)^T \end{pmatrix} \quad \forall s \in J$$

suggested by (55), has the unique zero $s = s^*$ in J . The bifurcation point x^* is said to be simple if δ changes sign at s^* .

When a continuation method is applied to approximate the path π , the process typically jumps over the bifurcation point. Accordingly, by monitoring the sign of δ , the bifurcation point can be detected. The computation of δ in the framework of the continuation method was sketched in Subsection 4.4. Of course, after passing a simple bifurcation point, it has to be taken into account that the two segments of the path before and after the point have opposite orientation. A scalar-equation solver, such as the Brent algorithm, can be applied for the explicit computation of the simple zero s^* of δ . There are also other approaches involving augmented nonlinear systems of equations that will not be addressed here.

By the definition of π , it follows from $\text{rank } DF(x^*) = m-1$ that the nullspace of $DF(x^*)$ is two-dimensional and contains the (nonzero) tangent vector $\pi'(s^*)$ of π at s^* . As the example (67) shows, once some unfolding \tilde{F} of F has been chosen, there is a vector $u \in \ker DF(x^*)$ that is linearly independent of $\pi'(s^*)$ and represents the direction of

a solution path of \tilde{F} branching-off from x^* . But, of course, without the unfolding no such information is available.

5.3 Higher-dimensional parameter spaces

Consider now a parameterized system (44) with a parameter space $\Lambda := \{0\} \times \mathbb{R}^d$ of dimension $d > 1$ defined by a C^r mapping F that is a submersion on its domain E . Hence, $M := F^{-1}(0)$ is a d -dimensional manifold.

For a given nonzero vector $b \in \{0\} \times \Lambda$, we introduce the augmented system

$$G(x, U) := \begin{pmatrix} F(x) \\ DF(x)U \\ U^T U \\ b^T U \end{pmatrix} = 0, \quad x \in E, \\ U \in L(\mathbb{R}^d, \mathbb{R}^n) \quad (74)$$

Any solution (x^*, U^*) of (74) satisfies $x^* \in M$, $U^* \in T_{x^*} M$, and $b \perp T_{x^*} M$. The last of these relations is equivalent to (66) and hence x^* is a foldpoint of M (with respect to Λ). In other words, systems of the form (74) can be used to compute foldpoints of M . Of course, the choice of b controls the foldpoints that can be obtained. In practice, attention is often focused on foldpoints with respect to a specific component of Λ in which case b is chosen as the canonical basis vector of \mathbb{R}^d corresponding to that component.

Since the dimension of the system (74) may become large and unwieldy, interest has focused on 'minimal augmentations' of the form

$$G(x) := \begin{pmatrix} F(x) \\ b^T u_1(x) \\ \vdots \\ b^T u_d(x) \end{pmatrix} = 0, \quad x \in E_0 \subset E \quad (75)$$

defined on some open set E_0 ; see, for example, Griewank and Reddien (1984). Here, $\{u_1(x), \dots, u_d(x)\}$ is for $x \in E_0 \cap M$ an orthonormal basis of $T_x M$, and, as before, $b \in \{0\} \times \Lambda$ is a suitably chosen vector. For computation, the basis vectors have to be sufficiently smooth functions of x and hence have to form a moving frame; see, for example, Rheinboldt (1988).

A different minimal augmentation for the computation of a simple foldpoint x^* of M with first singularity index 1 was developed by Dai and Rheinboldt (1990). For vectors $c^* \notin \text{rge } D_y F(x^*)$ and $v^* \notin \text{rge } D_y F(x^*)^T$, the linear system

$$\begin{pmatrix} D_y F(x)^T & v^* \\ (c^*)^T & 0 \end{pmatrix} \begin{pmatrix} z \\ \gamma \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0$$

is nonsingular for all $x \in \mathcal{M}$ in some neighborhood $\mathcal{U} \subset \mathbb{R}^d$ of x^* . Then Dai and Rheinboldt (1990) showed that all foldpoints of \mathcal{M} in $\mathcal{U} \cap \mathcal{M}$ are precisely the solutions of the augmented system

$$G(x) := \begin{pmatrix} F(x) \\ \gamma(x) \end{pmatrix} = 0$$

and that these foldpoints have first singularity index 1. Moreover, under certain nondegeneracy conditions on x^* and after shrinking \mathcal{U} if needed, the mapping G is a submersion on \mathcal{U} and hence the foldpoints of \mathcal{M} in that neighborhood form a $(d-1)$ -dimensional manifold. This allowed for the development of a locally convergent iterative process for the computation of a simplicial approximation of the foldpoint manifold. This was illustrated by an example concerning the roll-stability of airplanes.

The mapping F considered here may represent an unfolding of some other mapping F_0 involving fewer parameters, and therefore, the foldpoints of F may represent bifurcation points of F_0 . Then it is of interest to characterize and compute the corresponding bifurcation directions. For this, we follow here an approach developed by Rabier and Rheinboldt (1990) involving the second fundamental tensor of \mathcal{M} .

For the definition of this tensor, we refer to the textbook literature. In brief, V is a symmetric, vector-valued, 2-covariant tensor that defines, for $x \in \mathcal{M}$, a mapping

$$(u^1, u^2) \in T_x \mathcal{M} \longmapsto V_x(u^1, u^2) \in N_x \mathcal{M}$$

Let $Q(x)$ be the orthogonal projection of \mathbb{R}^d onto $N_x \mathcal{M}$. Then the diagonal terms of the tensor satisfy

$$V_x(u, u) := -Q(x)[DF(x)|_{T_x \mathcal{M}}]^{-1} D^2 F(x_0)(u, u), \quad u \in T_x \mathcal{M} \quad (76)$$

For computational purposes, this suffices since by the bilinearity

$$V_x(u^1, u^2) = \frac{1}{2} [V_x(u^1 + u^2, u^1 + u^2) - V_x(u^1, u^1) - V_x(u^2, u^2)], \quad u^1, u^2 \in T_x \mathcal{M}$$

A simple implementation of (76) by means of the QR-factorization of $DF(x)^T$ is included in the MANPACK package of Rheinboldt (1996). An algorithm for approximating $V_x(u, u)$ without requiring the second derivative of F was given by Rabier and Rheinboldt (1990).

Let x^* be a foldpoint of \mathcal{M} (with respect to Λ) with first singularity index $s_1 > 0$ and assume that $\{z^1, \dots, z^{s_1}\}$ is an orthonormal basis of N_0 . Then, subject to certain conditions

on x^* , Rabier and Rheinboldt (1990) showed that the bifurcation directions at x^* are the nonzero solutions $u \in T_{x^*} \mathcal{M}$ of the system of s_1 quadratic equations

$$(z^i)^T V_{x^*}(u, u) = 0, \quad i = 1, \dots, s_1 \quad (77)$$

Because of the obvious scale invariance, the solutions are here assumed to have norm 1; see, Rabier and Rheinboldt (1990) for several examples.

In bifurcation theory, the bifurcation directions are usually defined by means of a mapping generated in the Lyapunov-Schmidt reduction of the equation $F(x) = 0$. While both characterizations give equivalent results, the approach via the second fundamental tensor allows for all computations to be performed in the framework of the smooth manifold $F^{-1}(0)$ where no singularities hinder the work.

5.4 Further topics

As noted, the literature on bifurcation problems and their computational aspects is extensive and there are numerous major topics that cannot be considered here. We mention only a few important areas without entering into any detail.

Frequently, in computational mechanics, a nonlinear system of equations $F(y, \lambda) = 0$ arises as an equilibrium problem of some dynamical process, such as

$$\frac{dy}{dt} = F(y, \lambda) \quad (78)$$

In many such problems, the stability of the equilibria is highly important and hence it is of interest to obtain information about the stability behavior of (78). Now, bifurcation theory concerns the study of points where the qualitative structure of the solutions of (78) change as the parameter is varied. A change in the flow structure is often indicated by changes in the stability of the equilibria. For example, for (78) with $F(y, \lambda) := \lambda - y^2$, the parabolic path of equilibria is stable on one side of the foldpoint and unstable on the other. At a generic bifurcation point, there is typically an 'exchange of stability' between the equilibrium branches. For instance, in the case of (78) with the pitchfork function (67), the equilibrium branch along the negative λ axis gives up its stability to the two parabolic branches.

There are various other ways in which the qualitative structure of the flow of (78) can change. Important examples are the Hopf bifurcations. For instance, in the problem $\dot{r} = \lambda r - r^3$, $\dot{\theta} = -1$ (in polar coordinates), all solutions for $\lambda > 0$ spiral clockwise to the origin with increasing time. For $\lambda < 0$, the origin has become unstable and all solutions (except the unstable equilibrium at the origin) are spiraling into a periodic orbit of radius $\sqrt{\lambda}$.

For an introduction to dynamics and bifurcation, we refer to Hale and Koçak (1991). Clearly, (78) represents only the simplest dynamical system that has the zeros of F as its equilibria. In fact, in many mechanical problems, the dynamics arise in the form of a partial differential equation.

Symmetry is a natural phenomenon in many applications and often reflects some invariance properties of the problem. Frequently, a parameterized mapping $F: E \subset Y \times \Lambda \rightarrow \mathbb{R}^m$ turns out to be 'covariant' with respect to a transformation group Γ in the sense that

$$F(T_\gamma(y)y, T_\Lambda(\gamma)\lambda) = S(\gamma)F(y, \lambda) \quad \forall \gamma \in \Gamma, \quad (y, \lambda) \in E \quad (79)$$

where T_γ , T_Λ , and γ are group representations of Γ . For example, the unfolded pitchfork problem $F(y, \lambda, \mu) := \lambda y - y^3 - \mu$ is covariant under the Z_2 -symmetry $F(-y, \lambda, -\mu) = -F(y, \lambda, \mu)$. There is a large literature on symmetry behavior, and in particular, on symmetry breaking at bifurcation points of F . Group theoretical methods can be used effectively in numerical computations, for example, for detecting and evaluating bifurcation points. Symmetries also allow for the system to be restricted to certain subspaces defined by the groups under consideration. In addition, symmetries help in designing easy orderings of the computed results. The computational aspects of this topic were surveyed by Dellnitz and Werner (1989).

In practice, parameterized equations often arise in the form of boundary value problems for partial differential equations. In other words, the system $F(y, \lambda) = 0$ under consideration involves an operator $F: X \times Y \times \mathbb{R}^d \rightarrow Z$ between infinite-dimensional spaces X, Z , usually assumed to be Banach spaces. For the computation, this system has to be discretized, that is, a family $F_h: X \rightarrow Z$ of finite-dimensional approximations of F has to be constructed. Then the problem is to establish approximation theorems that guarantee the convergence of F_h to F as the discretization parameter $h \in \mathbb{R}$ tends to zero and that also provide error estimates. In addition, it is important to compare the solution sets of F and F_h and their structural properties, and to determine if singularities of F are approximated by singularities of F_h of the corresponding type. Results along this line were surveyed by Caloz and Rappaz (1997).

REFERENCES

- Allgower EL and Georg K. *Numerical Continuation Methods: An Introduction*. Springer-Verlag: New York, 1990.
- Allgower EL and Georg K. Numerical path following. In *Handbook of Numerical Analysis*, Ciarlet PG, Lions JL (eds), vol. V. Elsevier Science BV: Amsterdam, 1997; 3–207.

Allgower EL and Georg K. Piecewise linear methods for nonlinear equations and optimization. *J. Comput. Appl. Math.* 2000; 124:245–261.

Barnsley M. *Fractals Everywhere*. Academic Press: New York, 1988.

Berger MS. *Nonlinearity and Functional Analysis*. Academic Press: New York, 1977.

Brezinski C. *Projection Methods for Systems of Equations*. Elsevier Science B.V.: Amsterdam, 1997.

Burkardt J and Rheinboldt WC. A locally parameterized continuation process. *ACM Trans. Math. Softw.* 1983; 9:215–235.

Caloz G and Rappaz J. Numerical analysis for nonlinear and bifurcation problems. In *Handbook of Numerical Analysis*, Ciarlet PG, Lions JL (eds), vol. V. Elsevier Science B.V.: Amsterdam, 1997; 487–637.

Chow SN, Mallet-Paret J and Yorke JA. Finding zeros of maps: homotopy methods that are constructive with probability one. *Math. Comp.* 1978; 32:887–899.

Coleman TF and Moré JJ. Estimation of sparse Jacobian matrices and graph-coloring problems. *SIAM J. Numer. Anal.* 1983; 20:187–209.

Dai RX and Rheinboldt WC. On the computation of manifolds of foldpoints for parameter-dependent problems. *SIAM J. Numer. Anal.* 1990; 27:437–446.

Dellnitz M and Werner B. Computational methods for bifurcation problems with symmetries – with special attention to steady state and Hopf bifurcation points. *J. Comput. Appl. Math.* 1989; 26:97–123.

Denbo R, Eisenstat SC and Steihaug T. Inexact Newton methods. *SIAM J. Numer. Anal.* 1982; 19:400–408.

Dennis JE and Walker HF. Convergence theorems for least change secant update methods. *SIAM J. Numer. Anal.* 1981; 18:949–987.

Deuffhard P. A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with applications to multiple shooting. *Numer. Math.* 1974; 22:289–315.

Deuffhard P and Weiser M. Global inexact Newton multilevel FEM for nonlinear elliptic problems. *Multigrid Methods V*, (Stuttgart, 1996), *Lect. Notes Comput. Sci. Eng.*, vol. 3. Springer-Verlag: New York, 1998; 71–89.

Eisenstat SC and Walker HF. Globally convergent inexact Newton methods. *SIAM J. Opt.* 1994; 4:393–422.

Engelman M, Strang G and Bathe KJ. The application of quasi-Newton methods in fluid mechanics. *Int. J. Numer. Meth. Eng.* 1981; 17:707–718.

Floudas CA and Pardalos PM. *State of the Art in Global Optimization*. Kluwer Academic Publisher: Dordrecht, 1996.

Golub GH and van Loan CF. *Matrix Computations* (2nd edn). The Johns Hopkins University Press: Baltimore, 1989.

Golubitsky M and Schaeffer DG. *Singularities and Groups in Bifurcation Theory*, vol. I. Springer-Verlag: New York, 1985.

Griewank A. *Evaluating derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM Publications: Philadelphia, 2000.

Griewank A and Reddien GW. Characterization and computation of generalized turning points. *SIAM J. Numer. Anal.* 1984; 21:176–185.

- Hale JK and Koçak H. *Dynamics and Bifurcation*. Springer-Verlag: New York, 1991.
- Heitzinger W, Troch L and Valentia G. *Praxis nichtlinearer Gleichungen*, (in German). Carl Hanser Verlag: München, 1985.
- Henderson ME. Multiple parameter continuation: computing implicitly defined k-manifolds. *Int. J. Bifurc. Chaos* 2002; 12:451–476.
- Kearfott RB. *Rigorous Global Search: Continuous Problems*. Kluwer Academic Publisher: Dordrecht, 1996.
- Kelley CT. *Iterative Methods for Linear and Nonlinear Equations*, *Frontiers in Applied Mathematics*, vol. 16. SIAM Publications: Philadelphia, 1995.
- Melhem R and Rheinboldt WC. A comparison of methods for determining turning points of nonlinear equations. *Computing* 1982; 29:201–226.
- Moore G and Spence A. The calculation of turning points of nonlinear equations. *SIAM J. Numer. Anal.* 1980; 17:567–576.
- Nusse HE and Yorke JA. *Dynamics: Numerical Explorations*, Vol. 101 of *Appl. Math. Sciences*. Springer-Verlag: New York, 1994.
- Ortega JM and Rheinboldt WC. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press: 1970; 2nd edition, Vol. 30 of *Classics in Applied Mathematics*, SIAM Publications: Philadelphia, 2000; Russian translation 1976, Chinese translation 1982.
- Peitgen H-O and Richter PH. *The Beauty of Fractals*. Springer-Verlag: New York, 1986.
- Rabier PJ and Rheinboldt WC. On a computational method for the second fundamental tensor and its application to bifurcation problems. *Numer. Math.* 1990; 57:681–694.
- Rheinboldt WC. On measures of ill-conditioning for nonlinear equations. *Math. Comput.* 1976; 30:104–111.
- Rheinboldt WC. On the computation of multi-dimensional solution manifolds of parameterized equations. *Numer. Math.* 1988; 53:165–181.
- Rheinboldt WC. On the sensitivity of solutions of parameterized equations. *SIAM J. Numer. Anal.* 1993; 30:305–320.
- Rheinboldt WC. MANPACK: A set of algorithms for computations on implicitly defined manifolds. *Comput. Math. Appl.* 1996; 27:15–28.
- Rheinboldt WC. *Methods for Solving Systems of Nonlinear Equations* (2nd edn). *Regional Conf. Series in Appl. Math.*, vol. 70. SIAM Publications: Philadelphia, 1998.
- Watson LT, Sotomkina M, Melville RC, Morgan AP and Walker HF. Algorithm 777: HOMPACK90: a suite of FORTRAN 90 codes for globally convergent homotopy algorithms. *ACM Trans. Math. Softw.* 1997; 23:514–549.

Chapter 24

Adaptive Computational Methods for Parabolic Problems

K. Eriksson, C. Johnson and A. Logg

Chalmers University of Technology, Göteborg, Sweden

| | |
|--|-----|
| 1 What is a Parabolic Problem? | 675 |
| 2 Outline | 676 |
| 3 References to the Literature | 676 |
| 4 Introduction to Adaptive Methods for IVPs | 677 |
| 5 Examples of Stiff IVPs | 680 |
| 6 A Nonstiff IVP: The Lorenz System | 680 |
| 7 Explicit Time-stepping for Stiff IVPs | 683 |
| 8 Strong Stability Estimates for an Abstract Parabolic Model Problem | 686 |
| 9 Adaptive Space-Time Galerkin Methods for the Heat Equation | 689 |
| 10 A Priori and A Posteriori Error Estimates for the Heat Equation | 690 |
| 11 Adaptive Methods/Algorithms | 691 |
| 12 Reliability and Efficiency | 691 |
| 13 Strong Stability Estimates for the Heat Equation | 691 |
| 14 A Priori Error Estimates for the L_2 - and Elliptic Projections | 692 |
| 15 Proof of the A Priori Error Estimates | 693 |
| 16 Proof of the A Posteriori Error Estimates | 695 |
| 17 Extension to Systems of Convection-Diffusion-reaction Problems | 696 |
| 18 Examples of Reaction-Diffusion Problems | 696 |
| 19 Comparison with the Standard Approach to Time Step Control for ODEs | 699 |

| | |
|-----------------|-----|
| 20 Software | 702 |
| References | 702 |
| Further Reading | 702 |

The simpler a hypothesis is, the better it is. (Leibniz)

1 WHAT IS A PARABOLIC PROBLEM?

A common classification of partial differential equations uses the terms *elliptic*, *parabolic*, and *hyperbolic*, with the stationary Poisson equation being a prototype example of an elliptic problem, the time-dependent heat equation that of a parabolic problem, and the time-dependent wave equation being a hyperbolic problem. More generally, parabolic problems are often described, vaguely speaking, as 'diffusion-dominated', while hyperbolic problems are described as 'convection-dominated' in a setting of systems of convection-diffusion equations. Alternatively, the term 'stiff problems' is used to describe parabolic problems, with the term stiff referring to the characteristic presence of a range of time scales, varying from slow to fast with increasing damping.

In the context of computational methods for a general class of systems of time-dependent convection-diffusion-reaction equations, the notion of 'parabolicity' or 'stiffness' may be given a precise quantitative definition, which will be focal point of this presentation. We will define a system of convection-diffusion-reaction equations to be *parabolic* if computational solution is possible over a long time without error accumulation, or alternatively, if a certain *strong stability factor* $S_c(T)$, measuring error accumulation, is of

unit size independent of the length T in time of the simulation. More precisely, the error accumulation concerns the *Galerkin discretization error* in a *discontinuous Galerkin method* $dG(q)$ using piecewise polynomials of degree q with a resulting order of $2q + 1$. (The total discretization error may also contain a *quadrature error*, which typically accumulates at a linear rate in time for a parabolic problem.) This gives parabolicity a precise quantitative meaning with a direct connection to computational methods. A parabolic problem thus exhibits a feature of 'loss of memory' for Galerkin errors satisfying an orthogonality condition, which allows long-time integration without error accumulation. As shall be made explicit below, our definition of parabolicity through a certain stability factor is closely related to the definition of an *analytic semigroup*.

For a typical hyperbolic problem, the corresponding strong stability factor will grow linearly in time, while for more general initial value problems the growth may be polynomial or even exponential in time.

The solutions of parabolic systems, in general, vary considerably in space-time and from one component to the other with occasional *transients*, where derivatives are large. Efficient computational methods for parabolic problems thus require *adaptive* control of the mesh size in both space and time, or more general *multiadaptive* control with possibly different resolution in time for different components (see Chapter 4, Chapter 6, Chapter 7 of Volume 3).

2 OUTLINE

We first consider in Section 4 time-stepping methods for Initial Value Problems (IVPs) for systems of ordinary differential equations. We present an a posteriori error analysis exhibiting the characteristic feature of a parabolic problem of nonaccumulation of Galerkin errors in the setting of the backward Euler method (the discontinuous Galerkin method $dG(0)$), with piecewise constant (polynomial of order 0) approximation in time. The a posteriori error estimate involves the residual of the computed solution and stability factors/weights obtained by solving an associated dual linearized problem expressing in quantitative form the stability features of the IVP being solved. The a posteriori error estimate forms the basis of an adaptive method for time step control with the objective of controlling the Euclidean norm of the error uniformly in time or at selected time levels, or some other output quantity. The form of the a posteriori error estimate expresses the characteristic feature of a parabolic problem that the time step control is independent of the length in time of the simulation.

In Section 5, we compute stability factors for a couple of IVPs modeling chemical reactions and find that the strong stability factor $S_\epsilon(T)$ remains of unit size over a long time.

In Section 6, we contrast an IVP with exponentially growing stability factors: the Lorenz system.

The backward Euler method, or more generally the $dG(q)$ method, is implicit and requires the solution of a nonlinear system of equations at each time step. In Section 7, we study iterative fixed point-type solution strategies resembling explicit time-stepping methods. However, since explicit time-stepping for stiff problems is unstable unless the time step is smaller than the fastest time scale, which may be unnecessarily restrictive outside fast transients, we include a stabilization technique based on adaptively stabilizing the stiff system by taking a couple of small time steps when needed. We show efficiency gain factors compared to traditional explicit methods with the time step restriction indicated, of the order 10 to 100 or more depending on the problem. The need for explicit-type methods for parabolic problems avoiding forming Jacobians and solving associated linear systems of equations, is very apparent for the large systems of convection-diffusion-reaction equations arising in the modeling of chemical reactors with many reactants and reactions involved. The need for explicit time-stepping with the time step varying in both space and for different reactants, since here the discrete equations may be coupled over several time steps for some of the subdomains (or reactants), leading to very large systems of algebraic equations.

In Section 8, we prove the basic strong stability estimates for an abstract parabolic model problem and connect to the definition of an analytic semigroup.

In Sections 9 to 16, we present adaptive space-time Galerkin finite element methods for a model parabolic IVP, the heat equation, including a priori and a posteriori error estimates. The space-time Galerkin discretization method $cG(p)dG(q)$ is based on the continuous Galerkin method $cG(p)$ with piecewise polynomials of degree p in space, and the discontinuous Galerkin method $dG(q)$ with piecewise polynomials of degree q in time (for $q = 0, 1$). In Section 17, we discuss briefly the extension to convection-diffusion-reaction systems, and present computational results in Section 18.

3 REFERENCES TO THE LITERATURE

We have decided to give a coherent, concise presentation using duality-based, space-time Galerkin methods (with some key references), rather than to try to give a survey of the work in the entire field of parabolic or stiff problems (with a massive number of references). However, we believe that our presentation may be viewed as a summary

and condensation of much work prior to ours to which specific references can be found in the references given below. This work includes the pioneering work in the early 1970s initiated in Douglas and Dupont (1970) on Galerkin methods for parabolic problems and by Thomée during the 80s and 90s summarized in Thomée (2002) including many references to related work by Lusk, Rannacher, Wheeler, and many others.

It is of course also highly relevant to point to the world of stiff ordinary differential equations (stiff ode's) discovered by Dahlquist in the 50s, and explored by Deufelhard, Gear, Hairer, Lubich, Petzold, Wanner, and many others. As general references into this world, we give the books by Hairer and Wanner (1996) and Deufelhard and Bornemann (2002). In the concluding Section 19, we give our view of the relation between our methods for adaptive control of the global error and the methods for local error control commonly presented in the ode-world. Concerning Galerkin methods for time discretization we also refer to the early work on ode's by Delfour, Hager and Trochu (1981).

4 INTRODUCTION TO ADAPTIVE METHODS FOR IVPs

We now give a brief introduction to the general topic of *adaptive error control* for numerical time-stepping methods for initial value problems, with special reference to parabolic or stiff problems. In an *adaptive method*, the time steps are chosen automatically with the purpose of controlling the numerical error to stay within a given tolerance level. The adaptive method is based on an *a posteriori error estimate* involving the *residual* of the computed solution and results of auxiliary computations of *stability factors*, or more generally *stability weights*.

We consider an IVP of the form

$$\dot{u}(t) = f(u(t)) \quad \text{for } 0 < t \leq T, \quad u(0) = u^0 \quad (1)$$

where $f: \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a given differentiable function, $u^0 \in \mathbb{R}^d$ a given initial value, and $T > 0$ a given final time. For the computational solution of (1), we let $0 = t_0 < t_1 < \dots < t_{n-1} < t_n < \dots < t_N = T$ be an increasing sequence of discrete time steps with corresponding time intervals $I_n = (t_{n-1}, t_n]$ and time steps $k_n = t_n - t_{n-1}$, and consider the *backward Euler method*: Find $U(t_n)$ successively for $n = 0, 1, \dots, N$, according to the formula

$$U(t_n) = U(t_{n-1}) + k_n f(U(t_n)) \quad (2)$$

with $U(0) = u^0$. The backward Euler method is *implicit* in the sense that to compute the value $U(t_n)$ with $U(t_{n-1})$

already computed, we need to solve a system of equations. We will return to this aspect below.

We associate a *function* $U(t)$ defined on $[0, T]$ to the nodal values $U(t_n)$, $n = 0, 1, \dots, N$, as follows:

$$U(t) = U(t_n) \quad \text{for } t \in (t_{n-1}, t_n]$$

In other words, $U(t)$ is left-continuous piecewise constant on $[0, T]$ and takes the value $U(t_n)$ on I_n , and thus takes a jump from the limit from the left $U(t_{n-1}^-) = U(t_{n-1})$ to the limit from the right $U(t_{n-1}^+) = U(t_n)$ at the time level $t = t_{n-1}$. We can now write the backward Euler method in the form

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t)) \, dt$$

or equivalently

$$U(t_n) \cdot v = U(t_{n-1}) \cdot v + \int_{t_{n-1}}^{t_n} f(U(t)) \cdot v \, dt \quad (3)$$

for all $v \in \mathbb{R}^d$ with the dot signifying the scalar product in \mathbb{R}^d . This method is also referred to as $dG(0)$, the *discontinuous Galerkin method of order zero*, corresponding to approximating the exact solution $u(t)$ by a piecewise constant function $U(t)$ satisfying the *Galerkin orthogonality condition* (3).

The general $dG(q)$ method takes the form (3), with the restriction to each time interval I_n of the solution $U(t)$ and the test function v on each time interval I_n being a polynomial of degree q . The $dG(q)$ method also comes in a *multiadaptive form* with each component and the corresponding test function being piecewise polynomial with possibly different sequences of time steps for different components.

We shall now derive an *a posteriori error estimate* aiming at control of the scalar product of the error $e(T) = (u - U)(T)$ at final time T with a given vector ψ , where we assume that ψ is normalized so that $\|\psi\| = 1$. Here $\|\cdot\|$ denotes the Euclidean norm on \mathbb{R}^d . We introduce the following linearized dual problem running backward in time:

$$-\dot{\phi}(t) = A^T(t)\phi(t) \quad \text{for } 0 \leq t < T, \quad \phi(T) = \psi \quad (4)$$

with

$$A(t) = \int_0^1 f'(su(t) + (1-s)U(t)) \, ds$$

where $u(t)$ is the exact solution and $U(t)$ the approximate solution, f' is the Jacobian of f , and T denotes transpose.

We note that $f(u(t)) - f(U(t)) = A(t)(u(t) - U(t))$. We now start from the identity

$$e(T) \cdot \psi = e(T) \cdot \psi + \sum_{n=1}^N \int_{t_{n-1}}^{t_n} e \cdot (-\dot{\phi} - A^T \phi) dt$$

and integrate by parts on each subinterval (t_{n-1}, t_n) to get the error representation:

$$e(T) \cdot \psi = \sum_{n=1}^N \int_{t_{n-1}}^{t_n} (\dot{e} - Ae) \cdot \phi dt - \sum_{n=1}^N \langle U(t_n) - U(t_{n-1}), \phi(t_{n-1}) \rangle$$

where the last term results from the jumps of $U(t)$ at the nodes $t = t_{n-1}$. Since now u solves the differential equation $\dot{u} - f(u) = 0$, and $\dot{U} = 0$ on each time interval (t_{n-1}, t_n) , we have

$$\dot{e} - Ae = \dot{u} - f(u) - \dot{U} + f(U) = -\dot{U} + f(U) = f(U) \quad \text{on } (t_{n-1}, t_n)$$

It follows that

$$e(T) \cdot \psi = - \sum_{n=1}^N \langle U(t_n) - U(t_{n-1}), \phi(t_{n-1}) \rangle + \int_0^T f(U) \cdot \phi dt$$

Using the Galerkin orthogonality (3) with $v = \tilde{\phi}_n$, the mean value of ϕ over I_n , we get

$$e(T) \cdot \psi = - \sum_{n=1}^N \langle U(t_n) - U(t_{n-1}), (\phi(t_{n-1}) - \tilde{\phi}_n) \rangle + \sum_{n=1}^N \int_{t_{n-1}}^{t_n} f(U) \cdot (\phi - \tilde{\phi}_n) dt$$

Since now

$$\int_{t_{n-1}}^{t_n} f(U) \cdot (\phi - \tilde{\phi}_n) dt = 0$$

because $f(U(t))$ is constant on (t_{n-1}, t_n) , the error representation takes the form

$$e(T) \cdot \psi = - \sum_{n=1}^N \langle U(t_n) - U(t_{n-1}), (\phi(t_{n-1}) - \tilde{\phi}_n) \rangle$$

Finally, from the estimate

$$\|\phi(t_{n-1}) - \tilde{\phi}_n\| \leq \int_{t_{n-1}}^{t_n} \|\dot{\phi}(t)\| dt$$

we obtain the following *a posteriori* error estimate for the backward Euler or dG(0) method:

$$|e(T) \cdot \psi| \leq S_e(T, \psi) \max_{1 \leq n \leq N} \|U(t_n) - U(t_{n-1})\| \quad (5)$$

where the stability factor $S_e(T, \psi)$, recalling (4), is defined by

$$S_e(T, \psi) = \int_0^T \|\dot{\phi}(t)\| dt \quad (6)$$

Maximizing over ψ with $\|\psi\| = 1$, we obtain a posteriori control of the Euclidean norm of $e(T)$:

$$\|e(T)\| \leq S_e(T) \max_{1 \leq n \leq N} \|U(t_n) - U(t_{n-1})\| \quad (7)$$

with corresponding stability factor

$$S_e(T) = \max_{\|\psi\|=1} S_e(T, \psi) \quad (8)$$

Equivalently, we can write this estimate as

$$\|e(T)\| \leq S_e(T) \max_{0 \leq t \leq T} \|k(t) R(U(t))\| \quad (9)$$

where $k(t) = t_n - t_{n-1}$ for $t \in (t_{n-1}, t_n]$, and $R(U(t)) = (U(t_n) - U(t_{n-1}))/k_n = f(U(t_n))$ corresponds to the residual obtained by inserting the discrete solution into the differential equation (noting that $\dot{U}(t) = 0$ on each time interval).

We can, thus, also express the *a posteriori* error estimate (5) in the form

$$|e(T) \cdot \psi| \leq \int_0^T k(t) R(U(t)) \|\dot{\phi}(t)\| dt \quad (10)$$

where now the dual solution enters as a *weight* in a time integral involving the residual $R(U(t))$. Maximizing over $k(t) R(U(t))$ and integrating $\|\dot{\phi}(t)\|$ we obtain the original estimate (9).

We now define the IVP (1) to be *parabolic* if (up to possibly logarithmic factors) the stability factor $S_e(T)$ is of unit size for all T . We shall see that another typical feature of a parabolic problem is that the stability factor $S_e(T, \psi)$ varies little with the specific choice of normalized initial data ψ , which means that to compute $S_e(T) = \max_{\|\psi\|=1} S_e(T, \psi)$, we may drastically restrict the variation of ψ and solve the dual problem with only a few different initial data.

If we perturb f to \tilde{f} in the discretization with dG(q), for instance by approximating $f(U(t))$ by a polynomial connecting to quadrature in computing $\int_{t_n}^{t_{n+1}} f(U(t)) dt$, we obtain an additional contribution to the *a posteriori* error estimates of the form

$$S_q(T, \psi) \max_{0 \leq t \leq T} \|f(U(t)) - \tilde{f}(U(t))\|$$

or $S_q(T) \max_{0 \leq t \leq T} \|f(U(t)) - \tilde{f}(U(t))\|$, with corresponding stability factors defined by

$$S_q(T, \psi) = \int_0^T \|\dot{\phi}(t)\| dt$$

where ϕ solves the backward dual problem with $\phi(T) = \psi$, and $S_q(T) = \max_{\|\psi\|=1} S_q(T, \psi)$. In a parabolic problem, we may have $S_q(T) \sim T$, although $S_e(T) \sim 1$ for all $T > 0$. We note that dG(0) involves the time derivative $\dot{\phi}$, while $S_q(T)$ involves the dual ϕ itself.

Note that in dG(0) there is no need for quadrature in the present case of an autonomous IVP since then $f(U(t))$ is piecewise constant. However, in a corresponding nonautonomous problem of the form $\dot{u} = f(u(t), t)$ with f depending explicitly on t , quadrature may be needed also for dG(0).

The basic parabolic or stiff problem is a linear constant coefficient IVP of the form $\dot{u}(t) = f(u(t), t) = -Au(t) + f(t)$ for $0 < t \leq T$, $u(0) = u^0$, with A a constant positive semidefinite symmetric matrix with eigenvalues ranging from small to large positive. In this case, $f'(u) = -A$ with eigenvalues $\lambda \geq 0$ and corresponding solution components varying on time scales $1/\lambda$ ranging from very long (slow variation/decay if λ is small positive) to very short (fast variation/decay if λ is large positive). A solution to a typical stiff problem thus has a range of time scales varying from slow to fast. In this case, the dual problem takes the form $-\dot{\phi}(t) = -A\phi(t)$ for $0 \leq t < T$, and the strong stability estimate states that, independent of the distribution of the eigenvalues $\lambda \geq 0$ of A , we have

$$\int_0^T (T-t) \|\dot{\phi}(t)\|^2 dt \leq \frac{1}{4}$$

where we assume that $\|\phi(T)\| = 1$. From this we may derive that for $0 < \epsilon < T$,

$$\int_0^{T-\epsilon} \|\dot{\phi}(t)\| dt \leq \frac{1}{2} \left(\log \left(\frac{T}{\epsilon} \right) \right)^{1/2}$$

which up to a logarithmic factor states that $S_e(T) \sim 1$ for all $T > 0$. Further, the corresponding (weak) stability estimate states that $\|\dot{\phi}(t)\| \leq \|\psi\|$, from which it directly follows

that $S_q(T) \leq T$, as indicated. The (simple) proofs of the stability estimates are given below.

The stability factors $S_e(T, \psi)$ and $S_q(T, \psi)$ may be approximately computed *a posteriori* by replacing $A(t)$ in (4) with $f'(U(t))$, assuming $U(t)$ is sufficiently close to $u(t)$ for all t , and solving the corresponding backward dual problem numerically (e.g. using the dG(0) method). We may similarly compute approximations of $S_e(T)$ and $S_q(T)$ by varying ψ . By computing the stability factors, we get concrete evidence of the parabolicity of the underlying problem, which may be difficult (or impossible) to assess analytically *a priori*. Of course, there is also a gradual degeneracy of the parabolicity as the stability factor $S_e(T)$ increases.

The *a posteriori* error estimate (7) can be used as the basis for an adaptive time-stepping algorithm, controlling the size of the Galerkin discretization error, of the form: For $n = 1, 2, \dots, N$, choose k_n so that

$$\|U(t_n) - U(t_{n-1})\| \approx \frac{\text{TOL}}{S_e(T)}$$

for some tolerance $\text{TOL} > 0$. Recalling that the characteristic feature of a parabolic problem is that $S_e(T) \sim 1$ for all $T > 0$, this means that the time step control related to the Galerkin discretization error will be independent of the length of the time interval of the simulation. This means that long-time integration without error accumulation is possible, which may be interpreted as some kind of 'parabolic loss of memory'. We note again that this concerns the Galerkin error only, which has this special feature as a consequence of the Galerkin orthogonality. However, the quadrature error may accumulate in time typically at a linear rate, and so a long-time simulation may require more accurate quadrature than a simulation over a shorter interval.

Remark 1. We now present a very simple parabolic model problem, where we can directly see the basic feature of long-time integration without nonaccumulation of dG(0), which we just proved for a general parabolic problem. The model problem is simply $\dot{u}(t) = f(t)$, where f does not depend on $u(t)$. For this problem, which is certainly parabolic with our definition, dG(0) takes the form

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(t) dt \quad (11)$$

It follows that dG(0) (with exact quadrature) coincides with the exact solution at the discrete time levels t_n and thus there is no error accumulation at all over a long time. The reason is clearly the mean value property (11) of dG(0) expressing Galerkin orthogonality. On the other hand, if

we use quadrature to compute the integral in (11), then we may expect the quadrature error in general to accumulate (at a linear rate).

5 EXAMPLES OF STIFF IVPs

We have stated above that a parabolic or stiff initial value problem $\dot{u}(t) = f(u(t))$ for $0 < t \leq T$, $u(0) = u^0$, may be characterized by the fact that the stability factor $S_c(T)$ is of moderate (unit) size independent of $T > 0$, while the norm of the linearized operator $f'(u(t))$ may be large, corresponding to the presence of large negative eigenvalues. Such initial value problems are common in models of chemical reactions, with reactions on a range of time scales varying from slow to fast. Typical solutions include so-called *transients* where the fast reactions make the solution change quickly over a short (initial) time interval, after which the fast reactions are 'burned out' and the slow reactions make the solution change on a longer time scale. We now consider a set of test problems that we solve by the adaptive dG(0) method, including computation of the strong stability factor $S_c(T)$.

5.1 Model problem: $\dot{u} + Au(t) = f(t)$ with A positive symmetric semidefinite

As indicated, the basic example of a parabolic IVP takes the form $\dot{u} + Au(t) = f(t)$ for $0 < t \leq T$, $u(0) = u^0$, where A is a positive semidefinite square matrix. We consider here the case

$$A = \begin{pmatrix} -4.94 & 2.60 & 0.11 & 0.10 & 0.06 \\ 2.60 & -4.83 & 2.69 & 0.17 & 0.10 \\ 0.11 & 2.69 & -4.78 & 2.69 & 0.11 \\ 0.10 & 0.17 & 2.69 & -4.83 & 2.60 \\ 0.06 & 0.10 & 0.11 & 2.60 & -4.94 \end{pmatrix}$$

with eigenvalues $(0, -2.5, -5, -7.5, -9.33)$. In Figure 1, we plot the solution, the dual solution and the stability factor $S_c(T, \psi)$ as a function of T for a collection of different initial values $\psi(T) = \psi$. We note that the variation with ψ is rather small: about a factor 4. We also note the initial transient, both for the solution itself and for the dual problem runs backwards in time.

5.2 The Akzo-Nobel system of chemical reactions

We consider next the so-called Akzo-Nobel problem, which is a test problem for solvers of stiff ODEs modeling chemical reactions: Find the concentrations $u(t) =$

$(u_1(t), \dots, u_6(t))$ such that for $0 < t \leq T$,

$$\begin{cases} \dot{u}_1 = -2r_1 + r_2 - r_3 - r_4 \\ \dot{u}_2 = -0.5r_1 - r_4 - 0.5r_5 + F \\ \dot{u}_3 = r_1 - r_2 + r_3 \\ \dot{u}_4 = -r_2 + r_3 - 2r_4 \\ \dot{u}_5 = r_2 - r_3 + r_5 \\ \dot{u}_6 = -r_5 \end{cases} \quad (12)$$

where $F = 3.3 \cdot (0.9/737 - u_2)$ and the reaction rates are given by $r_1 = 18.7 \cdot u_1^4 / (u_2)$, $r_2 = 0.58 \cdot u_3 u_4$, $r_3 = 0.58/34.4 \cdot u_1 u_5$, $r_4 = 0.09 \cdot u_1 u_4^2$ and $r_5 = 0.42 \cdot u_5^2 / (u_2)$, with the initial condition $u^0 = (0.437, 0.00123, 0, 0, 0, 0.367)$. In Figure 2, we plot the solution, the dual solution and the stability factor $S_c(T)$ as a function of T . We note the initial transients in the concentrations and their long-time, very slow variation after the active phase of reaction. We also note that $S_c(T)$ initially grows to about 3.5 and then falls back to a value around 2. This is a typical behavior for reactive systems, where momentarily during the active phase of reaction the perturbation growth may be considerable, while over a long time the memory of that phase fades. On the other hand, $S_c(T)$ grows consistently, which shows that fading memory requires some mean value to be zero (Galerkin orthogonality). We present below more examples of this nature exhibiting features of parabolicity.

6 A NONSTIFF IVP: THE LORENZ SYSTEM

The Lorenz system presented in 1972 by meteorologist Edward Lorenz:

$$\begin{cases} \dot{u}_1 = -10u_1 + 10u_2 \\ \dot{u}_2 = 28u_1 - u_2 - u_1u_3 \\ \dot{u}_3 = -\frac{8}{3}u_3 + u_1u_2 \\ u(0) = u^0 \end{cases} \quad (13)$$

is an example of an IVP with exponentially growing stability factors reflecting a strong sensitivity to perturbations. Lorenz chose the model to illustrate perturbation sensitivity in meteorological models, making forecasts of daily weather virtually impossible over a period of more than a week. For the Lorenz system, accurate numerical solution using double precision beyond 50 units of time seems impossible. Evidently, the Lorenz system is not parabolic.

The system (13) has three equilibrium points \bar{u} with $f(\bar{u}) = 0$: $\bar{u} = (0, 0, 0)$ and $\bar{u} = (\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$. The equilibrium point $\bar{u} = (0, 0, 0)$ is unstable with the corresponding Jacobian $f'(\bar{u})$ having one positive (unstable)

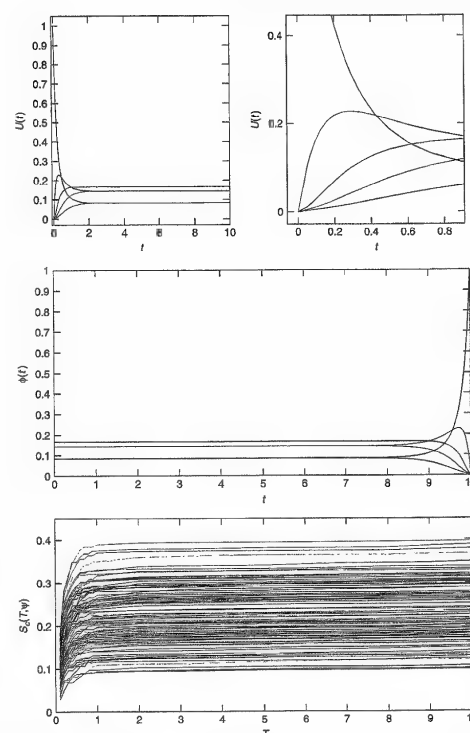


Figure 1. Symmetric IVP: solution, dual solution, and stability factors $S_c(T, \psi)$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecom>

eigenvalue and two negative (stable) eigenvalues. The equilibrium points $(\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$ are slightly unstable with the corresponding Jacobians having one negative (stable) eigenvalue and two eigenvalues with very small

positive real part (slightly unstable) and also an imaginary part. More precisely, the eigenvalues at the two nonzero equilibrium points are $\lambda_1 \approx -13.9$ and $\lambda_{2,3} \approx 0.0939 \pm 10.1i$.

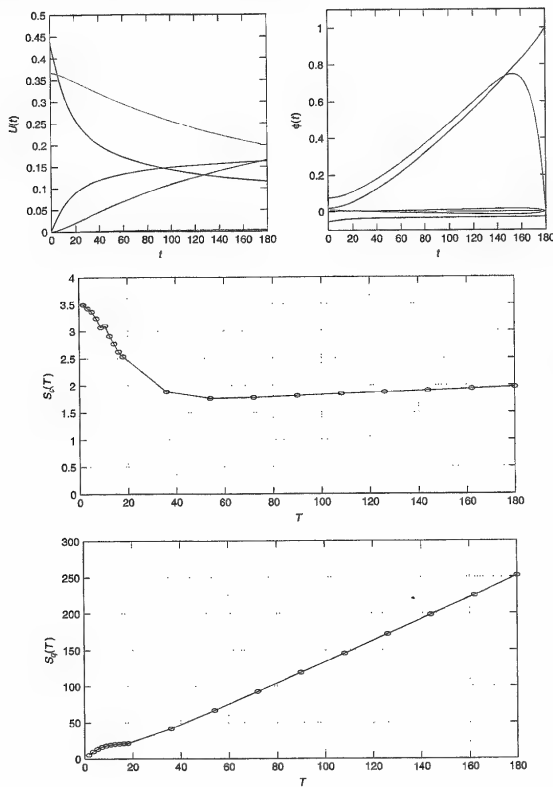


Figure 2. The Akzo-Nobel problem: solution, dual solution, stability factor $S_e(T, \psi)$, and stability factor $S_q(T, \psi)$. A color version of this image is available at <http://www.nrwi.interscience.wiley.com/ecm>

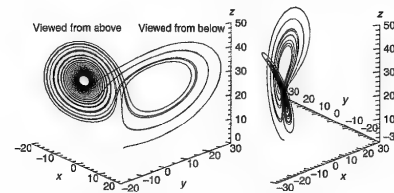


Figure 3. Two views of a numerical trajectory of the Lorenz system over the time interval $[0, 30]$.

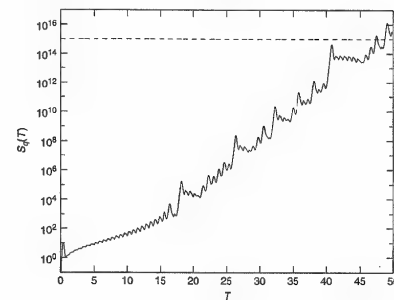


Figure 4. The growth of the stability factor $S_q(T)$ for the Lorenz problem. A color version of this image is available at <http://www.nrwi.interscience.wiley.com/ecm>

In Figure 3, we present two views of a solution $u(t)$ that starts at $u(0) = (1, 0, 0)$ computed to time 30 with an error tolerance of $TOL = 0.5$ using an adaptive IVP solver of the form presented above. The plotted trajectory is typical: it is kicked away from the unstable point $(0, 0, 0)$ and moves towards one of the nonzero equilibrium points. It then slowly orbits away from that point and at some time decides to cross over towards the other nonzero equilibrium point, again slowly orbiting away from that point and coming back again, orbiting out, crossing over, and so on. This pattern of some orbits around one nonzero equilibrium point followed by a transition to the other nonzero equilibrium point is repeated with a seemingly random number of revolutions around each nonzero equilibrium point.

In Figure 4, we plot the size of the stability factor $S_q(T)$ connected to quadrature errors as a function of final time T .

We notice that the stability factor takes an exponential leap every time the trajectory flips, while the growth is slower when the trajectory orbits one of the nonzero equilibrium points. The stability factor grows on an average as $10^{7/3}$, which sets the effective time limit of accurate computation to $T \approx 50$ computing in double precision with say 15 accurate digits.

7 EXPLICIT TIME-STEPPING FOR STIFF IVPs

The dG(0) method for the IVP $\dot{u} = f(u)$ takes the form

$$U(t_n) - k_n f(U(t_n)) = U(t_{n-1})$$

At each time step we have to solve an equation of the form $v - k_n f(v) = U(t_{n-1})$ with $U(t_{n-1})$ given. To this end, we may use a damped fixed-point iteration of the form

$$v^{(m)} = (1 - \alpha)v^{(m-1)} + \alpha(U(t_{n-1}) + k_n f(v^{(m-1)}))$$

with some suitable matrix α (or constant in the simplest case). Choosing $\alpha = I$ with only one iteration corresponds to the explicit Euler method. Convergence of the fixed-point iteration requires that

$$\|I - \alpha + k_n \alpha f'(v)\| < 1$$

for relevant values of v , which could force α to be small (e.g. in the stiff case with $f'(v)$ having large negative eigenvalues) and result in slow convergence. A simple choice is to take α to be a diagonal matrix with $\alpha_{ii} = 1/(1 - k_n f'_{ii}(v^{(m-1)}))$, corresponding to a diagonal approximation of Newton's method, with the hope that the number of iterations will be small.

We just learned that explicit time-stepping for stiff problems requires small time steps outside transients and thus may be inefficient. We shall now indicate a way to get around this limitation through a process of stabilization, where a large time step is accompanied by a couple of small time steps. The resulting method has similarities with the control system of a modern (unstable) jet fighter like the Swedish JAS Gripen, the flight of which is controlled by quick small flaps of a pair of small extra wings ahead of the main wings, or balancing a stick vertically on the finger tips if we want a more domestic application.

We shall now explain the basic (simple) idea of stabilization and present some examples as illustrations of fundamental aspects of adaptive IVP-solvers and stiff problems. Thus to start with, suppose we apply the explicit Euler method to the scalar problem

$$\begin{aligned} \dot{u}(t) + \lambda u(t) &= 0 \quad \text{for } 0 < t \leq T \\ u(0) &= u^0 \end{aligned} \quad (14)$$

with $\lambda > 0$ taking first a large time step K satisfying $K\lambda > 2$ and then m small time steps k satisfying $k\lambda < 2$, to get the method

$$U(t_n) = (1 - k\lambda)^m (1 - K\lambda) U(t_{n-1}) \quad (15)$$

altogether corresponding to a time step of size $k_n = K + mk$. Here K gives a large unstable time step with $|1 - K\lambda| > 1$ and k is a small time step with $|1 - k\lambda| < 1$. Defining the polynomial function $p(x) = (1 - kx)^m (1 - x)$, where $\theta = (k/K)$, we can write the method (15) in the form

$$U(t_n) = p(K\lambda) U(t_{n-1})$$

For stability, we need

$$|p(K\lambda)| \leq 1, \quad \text{that is } |1 - k\lambda|^m (K\lambda - 1) \leq 1$$

or

$$m \geq \frac{\log(K\lambda - 1)}{-\log|1 - k\lambda|} \approx 2 \log(K\lambda) \quad (16)$$

with $c = k\lambda \approx 1/2$ for definiteness.

We conclude that m may be quite small even if $K\lambda$ is large since the logarithm grows so slowly, and then only a small fraction of the total time (a small fraction of the time interval $[0, T]$) will be spent on stabilizing time-stepping with the small time steps k .

To measure the efficiency gain, we introduce

$$\alpha = \frac{1 + m}{K + km} \in \left(\frac{1}{K}, \frac{1}{k} \right)$$

which is the number of time steps per unit time interval with the stabilized explicit Euler method. By (16) we have

$$\alpha \approx \frac{1 + 2 \log(K\lambda)}{K + \log(K\lambda)/\lambda} \approx 2\lambda \frac{\log(K\lambda)}{K\lambda} \ll 2\lambda \quad (17)$$

for $K\lambda \gg 1$. On the other hand, the number of time steps per unit time interval for the standard explicit Euler method is

$$\alpha_0 = \frac{\lambda}{2} \quad (18)$$

with the maximum stable time step being $k_s = 2/\lambda$.

The cost reduction factor using the stabilized explicit Euler method would thus be

$$\frac{\alpha}{\alpha_0} \approx \frac{4 \log(K\lambda)}{K\lambda}$$

which can be quite significant for large values of $K\lambda$. For typical parabolic problems, $\dot{u} + Au(t) = 0$, the eigenvalues of A are distributed on the interval $[0, \lambda_{\max}]$, and for the damping to be efficient we need a slightly modified time step sequence. This is described in more detail in Eriksson, Johnson and Logg (2002).

We now present some examples using an adaptive cG(1) IVP-solver, where explicit fixed-point iteration (using only a couple of iterations) on each time interval is combined with stabilizing small time steps, as described for the explicit Euler method. In all problems, we note the initial transient, where the solution components change quickly, and the oscillating nature of the time step sequence outside the transient, with large time steps followed by some small stabilizing time steps.

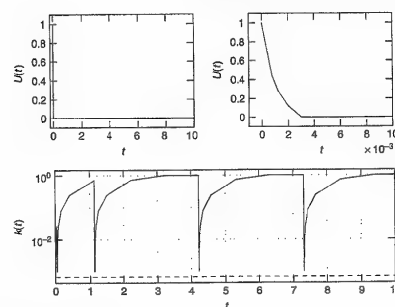


Figure 5. Solution and time step sequence for equation (14), $\alpha/\alpha_0 \approx 1/310$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecom>

Example We apply the indicated method to the scalar problem (14) with $u^0 = 1$ and $\lambda = 1000$ and display the result in Figure 5. The cost reduction factor in comparison to a standard explicit method is large: $\alpha/\alpha_0 \approx 1/310$.

Example We now consider the 2×2 diagonal system

$$\begin{aligned} \dot{u}(t) + \begin{pmatrix} 100 & 0 \\ 0 & 1000 \end{pmatrix} u(t) &= 0 \quad \text{for } 0 < t \leq T \\ u(0) &= u^0 \end{aligned} \quad (19)$$

with $u^0 = (1, 1)$. There are now two eigenmodes with large eigenvalues that need to be stabilized. The cost reduction factor is $\alpha/\alpha_0 \approx 1/104$ (see Figure 6).

Example We consider next the so-called HIREs problem ('High Irradiance RESponse') from plant physiology, which consists of the following eight equations:

$$\begin{aligned} \dot{u}_1 &= -1.71u_1 + 0.43u_2 + 8.32u_3 + 0.0007 \\ \dot{u}_2 &= 1.71u_1 - 8.75u_2 \\ \dot{u}_3 &= -10.03u_3 + 0.43u_4 + 0.035u_5 \\ \dot{u}_4 &= 8.32u_2 + 1.71u_3 - 1.12u_4 \\ \dot{u}_5 &= -1.74u_5 + 0.43u_6 + 0.43u_7 \\ \dot{u}_6 &= -280.0u_6u_8 + 0.69u_4 + 1.71u_5 - 0.43u_6 + 0.69u_7 \\ \dot{u}_7 &= 280.0u_6u_8 - 1.81u_7 \\ \dot{u}_8 &= -280.0u_6u_8 + 1.81u_7 \end{aligned} \quad (20)$$

together with the initial condition $u^0 = (1.0, 0, 0, 0, 0, 0, 0, 0, 0.0007)$. We present the solution and the time step sequence

in Figure 7. The cost is now $\alpha \approx 8$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/33$.

Example We consider again the Akzo-Nobel problem from above, integrating over the interval $[0, 180]$. We plot the solution and the time step sequence in Figure 8. Allowing a maximum time step of $k_{\max} = 1$ (chosen arbitrarily), the cost is $\alpha \approx 2$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/9$. The actual gain in a specific situation is determined by the quotient between the large time steps and the small damping time steps, as well as the number of small damping steps that are needed. In this case, the number of small damping steps is small, but the large time steps are not very large compared to the small damping steps. The gain is thus determined both by the stiff nature of the problem and the tolerance (or the size of the maximum allowed time step).

Example We consider now Van der Pol's equation:

$$\ddot{u} + \mu(u^2 - 1)\dot{u} + u = 0$$

which we write as

$$\begin{cases} \dot{u}_1 = u_2 \\ \dot{u}_2 = -\mu(u_1^2 - 1)u_2 - u_1 \end{cases} \quad (21)$$

We take $\mu = 1000$ and solve on the interval $[0, 10]$ with initial condition $u^0 = (2, 0)$. The cost is now $\alpha \approx 140$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/75$ (see Figure 9).

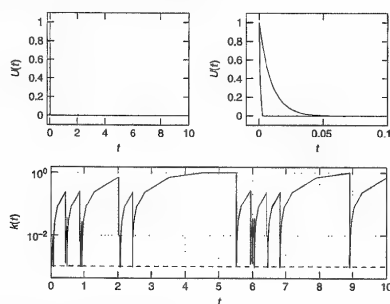


Figure 6. Solution and time step sequence for equation (19), $\alpha/\alpha_0 \approx 1/104$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

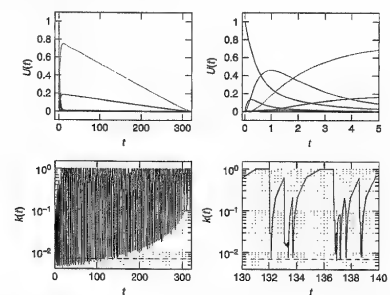


Figure 7. Solution and time step sequence for equation (20), $\alpha/\alpha_0 \approx 1/33$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

8 STRONG STABILITY ESTIMATES FOR AN ABSTRACT PARABOLIC MODEL PROBLEM

We consider an abstract parabolic model problem of the form: Find $w(t) \in H$ such that

$$\begin{cases} w_t(t) + Aw(t) = 0 & \text{for } 0 < t \leq T \\ w(0) = w^0 \end{cases} \quad (22)$$

where H is a vector space with inner product (\cdot, \cdot) and norm $\|\cdot\|$, A is a positive semidefinite symmetric linear operator defined on a subspace of H , that is, A is a linear transformation satisfying $(Aw, v) = (w, Av)$ and $(Av, v) \geq 0$ for all v, w in the domain of definition of A , and w^0 is the initial data. In the model problem of Section 4, $H = \mathbb{R}^d$ and A is a positive semidefinite symmetric $d \times d$ matrix. In the case of the heat equation, considered in the next section, $H = L_2(\Omega)$ and $-A = \Delta$

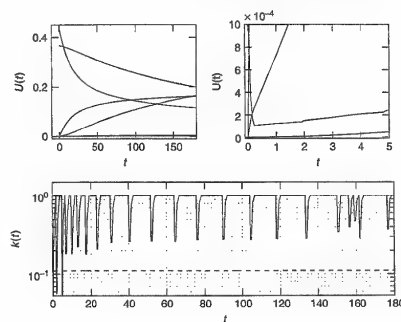


Figure 8. Solution and time step sequence for the Akzo-Nobel problem, $\alpha/\alpha_0 \approx 1/9$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

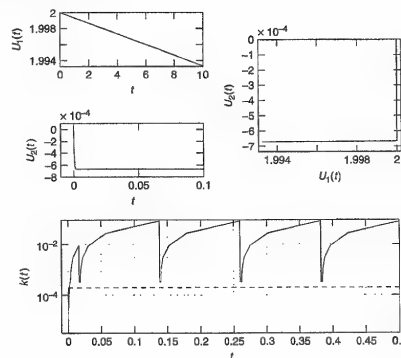


Figure 9. Solution and time step sequence for equation (21), $\alpha/\alpha_0 \approx 1/75$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecm>

(the Laplacian) with homogeneous Dirichlet boundary conditions.

It is important to notice that we do not require A to be positive definite, only positive semidefinite. In particular, A may have very small positive or zero eigenvalues, and our analysis does not depend on exponential decay of all eigenmodes in (22).

We now state and prove the basic strong stability estimates for the parabolic model problem (22), noting that the constants on the right-hand sides of the estimates are independent of the positive semidefinite symmetric operator A . It should be noted that the dual backward problem of (22), $-\dot{\phi} + A\phi = 0$, takes the form (22) with $w(t) = \phi(T-t)$.

Lemma 1. The solution w of (22) satisfies for $T > 0$,

$$\|w(T)\|^2 + 2 \int_0^T (Aw(t), w(t)) dt = \|w^0\|^2 \quad (23)$$

$$\int_0^T t \|Aw(t)\|^2 dt \leq \frac{1}{4} \|w^0\|^2 \quad (24)$$

$$\|Aw(T)\| \leq \frac{1}{\sqrt{2T}} \|w^0\| \quad (25)$$

Proof. Taking the inner product of $\dot{w}(t) + Aw(t) = 0$ with $w(t)$, we obtain

$$\frac{1}{2} \frac{d}{dt} \|w(t)\|^2 + (Aw(t), w(t)) = 0$$

from which (23) follows.

Next, taking the inner product of $\dot{w}(t) + Aw(t) = 0$ with $tAw(t)$ and using the fact that

$$(\dot{w}(t), tAw(t)) = \frac{1}{2} \frac{d}{dt} (t(Aw(t), w(t))) - \frac{1}{2} (Aw(t), w(t))$$

since A is symmetric, we find after integration that

$$\begin{aligned} & \frac{1}{2} T(Aw(T), w(T)) + \int_0^T t \|Aw(t)\|^2 dt \\ &= \frac{1}{2} \int_0^T (Aw(t), w(t)) dt \end{aligned}$$

from which (24) follows using (23) and the fact that $(Aw, w) \geq 0$.

Finally, taking the inner product in with $t^2 A^2 w(t)$, we obtain

$$\frac{1}{2} \frac{d}{dt} (t^2 \|Aw(t)\|^2) + t^2 (A^2 w(t), Aw(t)) = t \|Aw(t)\|^2$$

from which (25) follows after integration and using (24). \square

The estimates (23) to (25) express in somewhat different ways 'parabolic smoothing'; in particular, (25) expresses that the norm of the time derivative $\dot{w}(t)$, or equivalently $Aw(t)$, decreases (increases) like $1/t$ as t increases (decreases), which means that the solution becomes smoother with increasing time. We note a close relation between the two integrals

$$I_1 = \int_0^T \|\dot{w}(t)\| dt = \int_0^T \|Aw(t)\| dt$$

and

$$I_2 = \left(\int_0^T t \|\dot{w}(t)\|^2 dt \right)^{1/2} = \left(\int_0^T t \|Aw(t)\|^2 dt \right)^{1/2}$$

both measuring strong stability of (22), with I_1 through Cauchy's inequality being bounded by I_2 up to a logarithm:

$$\begin{aligned} \int_0^T \|Aw(t)\| dt &\leq \left(\int_0^T \frac{1}{t} dt \right)^{1/2} \left(\int_0^T t \|Aw(t)\|^2 dt \right)^{1/2} \\ &= \left(\log \left(\frac{T}{\epsilon} \right) \right)^{1/2} I_2 \end{aligned}$$

Remark 2. We now give an argument indicating that for the parabolic model problem (22), the stability factor $S_\epsilon(T, \psi)$ varies little with the specific choice of data ψ . We do this by noting that the quantity $S(w^0)$ defined by

$$S(w^0) = \left(\int_0^T t \|Aw(t)\|^2 dt \right)^{1/2}$$

where $w(t)$ solves (22) and varies little with the choice of initial data w^0 . To see this, we let $\{x_j\}$ be an orthonormal basis for H consisting of eigenfunctions of A with corresponding eigenvalues $\{\lambda_j\}$, which allows us to express the solution $w(t)$ in the form $\sum_j \exp(-\lambda_j t) w_j^0 x_j$ with $w_j^0 = (w^0, x_j)$. We may then write

$$\begin{aligned} (S(w^0))^2 &= \int_0^T t \sum_j \lambda_j^2 (w_j^0)^2 \exp(-2\lambda_j t) dt \\ &= \sum_j (w_j^0)^2 \int_0^T t \lambda_j^2 \exp(-2\lambda_j t) dt \end{aligned}$$

Now, the factor

$$\int_0^T t \lambda_j^2 \exp(-2\lambda_j t) dt = \int_0^{T\lambda_j} s \exp(-2s) ds$$

takes on almost the same value $\int_0^\infty s \exp(-2s) ds \approx 1/4$ for all j as soon as $T\lambda_j \geq 1$, that is, when λ_j is not very

small (since T is typically large). If we randomly choose the initial data w^0 , the chance of hitting an eigenfunction corresponding to a very small eigenvalue must be very small. We conclude that $S(w^0)$ varies little with w^0 . As just indicated, $S(w^0)$ is related to the integral $\int_0^T \|\dot{w}(t)\| dt$, which is the analog of the stability factor $S_\epsilon(T, \psi)$ for the dual problem. The bottom line is that $S_\epsilon(T, \psi)$ varies little with the choice of ψ .

Remark 3. The solution operator $\{E(t)\}_{t \geq 0}$ of (22), given by $E(t)w^0 = w(t)$, is said to define a uniformly bounded and analytic semigroup if there is a constant S such that the following estimates hold:

$$\begin{aligned} \|w(t)\| &\leq S \|w^0\| \\ \|Aw(t)\| &\leq \frac{S}{t} \|w^0\| \end{aligned} \quad (26)$$

for $t > 0$. We see that this definition directly couples to the stability estimates of Lemma 1, in which case the constant S is of unit size.

9 ADAPTIVE SPACE-TIME GALERKIN METHODS FOR THE HEAT EQUATION

We now move on to space-time Galerkin finite element methods for the model parabolic partial differential equation in the form of the heat equation: Find $u : \Omega \times I \rightarrow \mathbb{R}$ such that

$$\begin{cases} \dot{u} - \Delta u = f & \text{in } \Omega \times I \\ u = 0 & \text{on } \Gamma \times I \\ u(\cdot, 0) = u^0 & \text{in } \Omega \end{cases} \quad (27)$$

where Ω is a bounded domain in \mathbb{R}^d with boundary Γ , on which we have posed homogeneous Dirichlet boundary conditions, u^0 is a given initial temperature, f a heat source, and $I = (0, T]$ a given time interval.

For the discretization of the heat equation in space and time, we use the cG(p)dG(q) method based on a tensor product space-time discretization with continuous piecewise polynomial approximation of degree $p \geq 1$ in space and discontinuous piecewise polynomial approximation of degree $q \geq 0$ in time, giving a method which is accurate of order $p+1$ in space and of order $2q+1$ in time. The discontinuous Galerkin dG(q) method used for the time discretization reduces to the *subdiagonal Padé method* for homogeneous constant coefficient problems and in general, together with quadrature for the evaluation of the integral in time, corresponds to an *implicit Runge-Kutta method*. For the discretization in space, we use the standard conforming continuous Galerkin cG(p) method. The cG(p)dG(q)

method has maximal flexibility and allows the space and time steps to vary in both space and time. We design and analyze reliable and efficient adaptive algorithms for global error control in $L_\infty(L_2(\Omega))$ (maximum in time and L_2 in space), with possible extensions to $L_r(L_r(\Omega))$ with $1 \leq r, s \leq \infty$.

The cG(p)dG(q) method is based on a partition in time $0 = t_0 < t_1 < \dots < t_N = T$ of the interval $(0, T]$ into time intervals $I_n = (t_{n-1}, t_n]$ of length $k_n = t_n - t_{n-1}$ with associated finite element spaces $S_n \subset H_0^1(\Omega)$ consisting of piecewise polynomials of degree p on a triangulation $\mathcal{T}_n = \{K\}$ of Ω into elements K with local mesh size given by a function $h_n(x)$. We define

$$V_n = \left\{ v : v = \sum_{j=0}^q t^j v_j, v_j \in S_n \right\}$$

and

$$V = \{v : v|_{I_n} \in V_n, n = 1, \dots, N\}$$

We thus define V to be the set of functions $v : \Omega \times I \rightarrow \mathbb{R}$ such that the restriction of $v(x, t)$ to each time interval I_n is polynomial in t with coefficients in S_n . The cG(p)dG(q) method for (27) now reads: Find $U \in V$ such that for $n = 1, 2, \dots, N$,

$$\begin{aligned} & \int_{I_n} ((U, v) + (\nabla U, \nabla v)) dt + ((U|_{I_{n-1}}, v_{n-1}^+) \\ &= \int_{I_n} (f, v) dt \quad \forall v \in V_n \end{aligned} \quad (28)$$

where $[w]_n = w(t_n^+) - w(t_n^-)$, $w_n^{+(-)} = \lim_{t \rightarrow 0^{+(-)} w(t_n + s)$, $U_0^- = u^0$, and (\cdot, \cdot) denotes the $L_2(\Omega)$ or $[L_2(\Omega)]^d$ inner product. Note that we allow the space discretizations to change with time from one space-time slab $\Omega \times I_n$ to the next.

For $q = 0$, the scheme (28) reduces to the following variant of the backward Euler scheme:

$$U_n - k_n \Delta_n U_n = P_n U_{n-1} + \int_{I_n} P_n f dt \quad (29)$$

where $U_n \equiv U|_{I_n}$, $\Delta_n : S_n \rightarrow S_n$ is the discrete Laplacian on S_n defined by $(-\Delta_n v, w) = (\nabla v, \nabla w)$ for all $w \in S_n$, and P_n is the L_2 -projection onto S_n defined by $(P_n v, w) = (v, w)$ for all $w \in S_n$.

Alternatively, (29) may be written (with $f \equiv 0$) in matrix form as

$$M_n \xi_n + k_n A_n \xi_n = M_n \xi_{n-1}$$

where M_n and A_n are mass and stiffness matrices related to a nodal basis for S_n , ξ_n is the corresponding vector of nodal

values for U_n , and ξ_{n-1} is the vector of nodal values for $P_n U_{n-1}$. Evidently, we have to solve a system of equations with system matrix $M_n + k_n A_n$ to compute ξ_n .

Remark 4. Note that in the discretization (28), the space and time steps may vary in time and that the space discretization may be variable also in space, whereas the time steps k_n are kept constant in space. Clearly, optimal mesh design requires the time steps to be variable also in space. Now, it is easy to extend the method (28) to admit time steps that are variable in space simply by defining

$$V_n = \left\{ v : v|_{I_n} = \sum_j c_j(t) v_j, v_j \in S_n \right\}$$

where now the coefficients $c_j(t)$ are piecewise polynomial of degree q in t without continuity requirements on partitions of I_n which may vary with j . The discrete functions may now be discontinuous in time also inside the space-time slab $\Omega \times I_n$, and the degree q may vary over components and subintervals. The cG(p)dG(q) method again takes the form (28), with the term $(U)_{n-1}^{(q)}$ replaced by a sum over all jumps in time of U in $\Omega \times [t_{n-1}, t_n]$. Adaptive methods in this generality, so-called *multidaptive* methods, are proposed and analyzed in detail for systems of ordinary differential equations in Logg (2001a,b).

10 A PRIORI AND A POSTERIORI ERROR ESTIMATES FOR THE HEAT EQUATION

In this section, we state a priori and a posteriori error estimates for the cG(p)dG(q) method (28) in the case $p = 1$ and $q = 0, 1$ and give the proofs below. A couple of technical assumptions on the space-mesh function $h_n(x)$ and time steps k_n are needed: We assume that each triangulation T_n with associated mesh size h_n satisfies, with $h_{n,K}$ equal to the diameter and $m_{n,K}$ the volume of $K \in T_n$,

$$c_1 h_{n,K}^2 \leq m_{n,K} \quad \forall K \in T_n \quad (30)$$

$$c_2 h_{n,K} \leq h_n(x) \leq h_{n,K} \quad \forall x \in K \quad \forall K \in T_n \quad (31)$$

$$|\nabla h_n(x)| \leq \mu \quad \forall x \in \Omega \quad (32)$$

for some positive constants c_1, c_2 , and μ . The constant μ will be assumed to be small enough in the a priori error estimates (but not in the a posteriori error estimates). We further assume that there are positive constants c_3, c_4 , and

γ such that for all n we have

$$k_n \leq c_3 k_{n+1} \quad (33)$$

$$c_4 h_n(x) \leq h_{n+1}(x) \leq \frac{1}{c_4} h_n(x) \quad \forall x \in \Omega \quad (34)$$

$$\tilde{h}_n^2 \leq \gamma k_n \quad \text{or} \quad S_n \subset S_{n-1} \quad (35)$$

where $\tilde{h}_n = \max_{x \in \Omega} h_n(x)$. Furthermore, we assume for simplicity that Ω is convex, so that the following *elliptic regularity* estimate holds: $\|D^2 v\| \leq \|\Delta v\|$ for all functions v vanishing on Γ . Here $(D^2 v)^2 = \sum_{i,j} v_{,ij}^2$, where $v_{,ij}$ is the second partial derivative of v with respect to x_i and x_j , and $\|\cdot\|$ denotes the $L_2(\Omega)$ -norm. With these assumptions, we have the following a priori error estimates:

Theorem 1. If μ and γ are sufficiently small, then there is a constant C depending only on the constants c_i , $i = 1, 2, 3, 4$, such that for u the solution of (27) and U that of (28), we have for $p = 1$ and $q = 0, 1$,

$$\|u - U\|_{L_n} \leq CL_n \max_{1 \leq m \leq n} E_{qm}(u), \quad n = 1, \dots, N \quad (36)$$

and for $q = 1$,

$$\|u(t_n) - U(t_n)\| \leq CL_n \max_{1 \leq m \leq n} E_{2m}(u), \quad n = 1, \dots, N \quad (37)$$

where $L_n = (\log(t_n/k_n) + 1)^{1/2}$, $E_{qm}(u) = \min_{1 \leq q \leq m} k_m^q \|u_{,ij}^{(q)}\|_{L_n} + \|h_m^2 D^2 u\|_{L_n}$, $q = 0, 1, 2$ with $u_{,ij}^{(1)} = \dot{u}$, $u_{,ij}^{(2)} = \ddot{u}$, $u_{,ij}^{(3)} = \Delta \ddot{u}$ and $\|w\|_{L_n} = \max_{t \in I_n} \|w(t)\|$.

These estimates state that the discontinuous Galerkin method (28) is of order $q + 1$ globally in time and of order $2q + 1$ at the discrete time levels t_n for $q = 0, 1$, and is second order in space. In particular, the estimate (36) is *optimal* compared to interpolation with piecewise polynomials of order $q = 0, 1$ in time and piecewise linears in space, up to the logarithmic factor L_n . The third order accuracy in time at the discrete time levels for $q = 1$ reflects a *superconvergence* feature of the dG(q) method.

The a posteriori error estimates for (28) take the form:

Theorem 2. If u is the solution of (27) and U that of (28) with $p = 1$, then we have for $q = 0$,

$$\|u(t_n) - U(t_n)\| \leq \max_{1 \leq m \leq n} \mathcal{E}_{0m}(U), \quad n = 1, \dots, N \quad (38)$$

and for $q = 1$,

$$\|u(t_n) - U(t_n)\| \leq \max_{1 \leq m \leq n} \mathcal{E}_{2m}(U), \quad n = 1, \dots, N \quad (39)$$

where

$$\mathcal{E}_{0m}(U) = \gamma_1 \|h_m^2 R(U)\|_{L_n} + \gamma_2 \left\| \frac{h_m^2 [U]_{m-1}}{k_m} \right\|_{L_n} + \gamma_3 \|k_m R_{0k}(U)\|_{L_n}$$

$$\mathcal{E}_{2m}(U) = \gamma_1 \|h_m^2 R(U)\|_{L_n} + \gamma_2 \left\| \frac{h_m^2 [U]_{m-1}}{k_m} \right\|_{L_n} + \gamma_4 \|k_m^3 R_{1k}(U)\|_{L_n}$$

and

$$R(U) = |f| + D_m^2 U$$

$$R_{0k}(U) = |f| + \frac{|[U]_{m-1}|}{k_m}$$

$$R_{1k}(U) = |f_{,i}| + \frac{|\Delta_m P_m [U]_{m-1}|}{k_m^2}$$

on $\Omega \times I_n$. A star indicates that the corresponding term is present only if S_{m-1} is not a subset of S_m . Further, $\gamma_i = L_n C_i$, where the C_i are constants related to approximation by piecewise constant or linear functions. Finally, $D_m^2 U$ on a space element $K \in T_n$ is the modulus of the maximal jump in normal derivative of U across an edge of K divided by the diameter of K .

Remark 5. The term $|f|$ in $R(U)$ may be replaced by $|h_m^2 D^2 f|$. Similarly, the term $|f|$ in R_{0k} may be replaced with $k|f|$ and $|f|$ in $R_{1k}(U)$ by $k|\Delta f|$.

The a posteriori error estimates are sharp in the sense that the quantities on the right-hand sides can be bounded by the corresponding right-hand sides in the (optimal) a priori error estimates. Therefore, the a posteriori error estimates may be used as a basis for efficient adaptive algorithms, as we indicate below.

11 ADAPTIVE METHODS/ALGORITHMS

An adaptive method for the heat equation addresses the following problem: For a given tolerance $\text{TOL} > 0$, find a discretization in space and time $S_{hk} = \{(T_n, k_n)_{n \geq 1}\}$, such that

$$(1) \|u(t_n) - U(t_n)\| \leq \text{TOL} \quad \text{for } n = 1, 2, \dots$$

$$(2) S_{hk} \text{ is optimal, in the sense that the number of degrees of freedom is minimal} \quad (40)$$

We approach this problem using the a posteriori estimates (38) and (39) in an adaptive method of the form: Find S_{hk}

such that for $n = 1, 2, \dots$,

$$\mathcal{E}_{0k}(U) \leq \text{TOL}, \quad \text{if } q = 0$$

$$\mathcal{E}_{2k}(U) \leq \text{TOL}, \quad \text{if } q = 1$$

the number of degrees of freedom of

$$S_{hk} \text{ is minimal} \quad (41)$$

To solve this problem, we use an *adaptive algorithm* for choosing S_{hk} based on *equidistribution* of the form: For each $n = 1, 2, \dots$, with T_{n0} a given initial space mesh and k_{n0} an initial time step, determine triangulations T_{nj} with N_j elements of size $h_{nj}(x)$, time steps k_{nj} , and corresponding approximate solutions U_{nj} defined on $I_{nj} = (t_{n-1}, t_{n-1} + k_{nj})$, such that for $j = 0, 1, \dots, \tilde{n} - 1$,

$$\gamma_1 \max_{t \in I_{nj}} \|h_{nj}^2 R(U_{nj})\|_{L_2(K)} + \gamma_2 \left\| \frac{h_{nj}^2 [U]_{n-1,j}}{k_{nj}} \right\|_{L_2(K)} \leq \frac{\text{TOL}}{2\sqrt{N_j}} \quad \forall K \in T_{nj}$$

$$k_{n,j+1} \gamma_3 \|R_{0k}(U_{nj})\|_{L_{nj}} \leq \frac{\text{TOL}}{2}, \quad \text{if } q = 0$$

$$k_{n,j+1}^3 \gamma_4 \|R_{1k}(U_{nj})\|_{L_{nj}} \leq \frac{\text{TOL}}{2}, \quad \text{if } q = 1 \quad (42)$$

that is, we determine iteratively each new time step $k_n = k_{n\tilde{n}}$ and triangulation $T_n = T_{n\tilde{n}}$. The number of trials \tilde{n} is the smallest integer j such that (41) holds with U replaced by U_{nj} , and the parameter $\theta \sim 1$ is chosen so that \tilde{n} is small.

12 RELIABILITY AND EFFICIENCY

By the a posteriori estimates (38) and (39), it follows that the adaptive method (41) is *reliable* in the sense that if (41) holds, then the error control (40) is guaranteed. The *efficiency* of (41) follows from the fact that the right-hand sides of the a posteriori error estimates may be bounded by the corresponding right-hand sides in the (optimal) a priori error estimates.

13 STRONG STABILITY ESTIMATES FOR THE HEAT EQUATION

We now state the fundamental strong stability results for the continuous and discrete problems to be used in the proofs of the a priori and a posteriori error estimates. Analogous to Section 8, we consider the problem $\psi - \Delta w = 0$, where $w(t) = \phi(T - t)$ is the backward dual solution with time reversed.

The proof of Lemma 2 is similar to that of Lemma 1, multiplying by $w(t)$, $-t\Delta w(t)$, and $t^2\Delta^2 w(t)$. The proof of Lemma 3 is also analogous: For $q = 0$, we multiply (29) by W_n and $t_n A_n W_n$, noting that if $S_{n-1} \subset S_n$ (corresponding to coarsening in the time direction of the primal problem $u - \Delta u = f$), then $P_n W_{n-1} = W_{n-1}$, $(A_n W_n, W_{n-1}) = (W_n, A_n W_{n-1})$ and $(A_n W_{n-1}, W_{n-1}) = (A_{n-1} W_{n-1}, W_{n-1})$. The proof for $q = 1$ is similar.

Lemma 2. Let w be the solution of (27) with $f \equiv 0$. Then for $T > 0$,

$$\|w(T)\|^2 + 2 \int_0^T \|\nabla w(t)\|^2 dt = \|w^0\|^2 \quad (43)$$

$$\int_0^T t \|\dot{w}(t)\|^2 + \|\Delta w(t)\|^2 dt \leq \frac{1}{2} \|w^0\|^2 \quad (44)$$

$$\|\Delta w(T)\| \leq \frac{1}{\sqrt{2T}} \|w^0\| \quad (45)$$

Lemma 3. There is a constant C , such that if $S_{n-1} \subset S_n$ for $n = 1, 2, \dots, N$, and W is the solution of (28) with $f = 0$, then for $T = t_N > 0$,

$$\|W_N\|^2 + 2 \int_0^T \|\nabla W\|^2 dt + \sum_{n=1}^N \|W_{n-1}\|^2 = \|w^0\|^2 \quad (46)$$

$$\sum_{n=1}^N t_n \int_{I_n} \{\|\dot{W}\|^2 + \|\Delta_n W\|^2\} dt + \sum_{n=1}^N t_n \frac{\|W_{n-1}\|^2}{k_n} \leq C \|w^0\|^2 \quad (47)$$

and

$$\sum_{n=1}^N \int_{I_n} \{\|\dot{W}\| + \|\Delta_n W\|\} dt + \sum_{n=1}^N \|W_{n-1}\| \leq C \left(\log \frac{t_N}{k_1} + 1 \right)^{1/2} \|w^0\| \quad (48)$$

14 A PRIORI ERROR ESTIMATES FOR THE L_2 - AND ELLIPTIC PROJECTIONS

We shall use the following a priori error estimate for the L_2 -projection $P_n : L_2(\Omega) \rightarrow S_n$ defined by $(w - P_n w, v) = 0$ for all $v \in S_n$. This estimate follows from the fact that P_n is very close to the nodal interpolation operator J_n into S_n ,

defined by $J_n w = w$ at the nodes of T_n if w is smooth (and $J_n w = J_n \tilde{w}$ if $w \in H^1(\Omega)$, where \tilde{w} is a locally regularized approximation of w).

Lemma 4. If μ in (32) is sufficiently small, then there is a positive constant C such that for all $w \in H_0^1(\Omega) \cap H^2(\Omega)$,

$$|(f, w - P_n w) - (\nabla U, \nabla(w - P_n w))| \leq C \|h_n^2 R_n(U)\| \|D^2 w\| \quad (49)$$

where $R_n(U) = |f| + D_n^2 U$.

We shall also need the following a priori error estimate for the elliptic projection $\pi_n : H_0^1(\Omega) \rightarrow S_n$ defined by

$$(\nabla(w - \pi_n w), \nabla v) = 0 \quad \forall v \in S_n \quad (50)$$

Lemma 5. If μ in (32) is sufficiently small, then there is a positive constant C such that for all $w \in H^2(\Omega) \cap H_0^1(\Omega)$,

$$\|w - \pi_n w\| \leq C \|h_n^2 D^2 w\| \quad (51)$$

Proof. We shall first prove that with $e = w - \pi_n w$, we have $\|e\| \leq C \|h_n \nabla e\|$. For this purpose, we let ϕ be the solution of the continuous dual problem $-\Delta \phi = e$ in Ω with $\phi = 0$ on Γ , and note that by integration by parts, the Galerkin orthogonality (50), a standard estimate for the interpolation error $u - J_n u$, together with elliptic regularity, we have

$$\begin{aligned} \|e\|^2 &= (e, -\Delta \phi) = (\nabla e, \nabla \phi) = (\nabla e, \nabla(\phi - J_n \phi)) \\ &= |h_n \nabla e| \|h_n^{-1} \nabla(\phi - J_n \phi)\| \\ &\leq C \|h_n \nabla e\| \|D^2 \phi\| \leq C \|h_n \nabla e\| \|e\| \end{aligned}$$

which proves the desired estimate. Next, to prove that $\|h_n \nabla e\| \leq C \|h_n^2 D^2 w\|$, we note that since $\pi_n J_n u = J_n u$, we have

$$\begin{aligned} \|h_n \nabla e\| &\leq \|h_n \nabla(w - J_n w)\| + \|h_n \nabla \pi_n(w - J_n w)\| \\ &\leq C \|h_n \nabla(w - J_n w)\| \leq C \|h_n^2 D^2 w\|_2 \end{aligned}$$

where we used stability of the elliptic projection π_n in the form

$$\|h_n \nabla \pi_n v\| \leq C \|h_n \nabla v\| \quad \forall v \in H_0^1(\Omega)$$

which is a weighted analog of the basic property of the elliptic projection $\|\nabla \pi_n v\| \leq \|\nabla v\|$ for all $v \in H_0^1(\Omega)$. For the proof of the weighted analog, we need the mesh size

not to vary too quickly, expressed in the assumption that μ is small. \square

15 PROOF OF THE A PRIORI ERROR ESTIMATES

In this section, we give the proof of the a priori estimates, including (36) and (37). For simplicity, we shall assume that $S_n \subset S_{n-1}$, corresponding to a situation where the solution gets smoother with increasing time. The proof is naturally divided into the following steps, indicating the overall structure of the argument:

1. an error representation formula using duality;
2. strong stability of the discrete dual problem;
3. choice of interpolant and proof of (36); and
4. choice of interpolant and proof of (37).

15.1 An error representation formula using duality

Given a discrete time level $t_N > 0$, we write the discrete set of equations (28) determining the discrete solution $U \in V$ up to time t_N in compact form as

$$A(U, v) = (u^0, v_0^0) + (f, v)_I \quad \forall v \in V \quad (52)$$

where

$$\begin{aligned} A(w, v) &= \sum_{n=1}^N [(w, v)_n + (\nabla w, \nabla v)_n] + (w_0^0, v_0^0) \\ &\quad + \sum_{n=2}^N [(w)_{n-1}, v_{n-1}^+] \end{aligned}$$

$(v, w)_n = \int_{I_n} (v, w) dt$ and $I = (0, T]$. The error $e \equiv u - U$ satisfies the Galerkin orthogonality

$$A(e, v) = 0 \quad \forall v \in V \quad (53)$$

which follows from the fact that (52) is satisfied also by the exact solution u of (27). Let the discrete dual solution $\Phi \in V$ now be defined by

$$A(v, \Phi) = (v_N^-, e_N) \quad \forall v \in V \quad (54)$$

where $e_N = u(t_N) - U(t_N)$ is the error at final time t_N , corresponding to control of the $L_2(\Omega)$ -norm of e_N . We note

that Φ is a discrete cG(p)dG(q)-solution of the continuous dual problem

$$\begin{aligned} -\dot{\Phi} - \Delta \Phi &= 0 \quad \text{in } \Omega \times [0, T) \\ \Phi &= 0 \quad \text{on } \Gamma \times [0, T) \end{aligned} \quad (55)$$

with initial data $\Phi(T) = e_N$. This follows from the fact that the bilinear form $A(\cdot, \cdot)$, after time integration by parts, can also be written as

$$\begin{aligned} A(w, v) &= \sum_{n=1}^N [(w, -\dot{v})_n + (\nabla w, \nabla v)_n] + \sum_{n=1}^{N-1} (w_n^-, -[v]_n) \\ &\quad + (w_N, v_N) \end{aligned} \quad (56)$$

In view of (54) and (53), we have for any $v \in V$,

$$\begin{aligned} \|e_N\|^2 &= (u_N - v_N^-, e_N) + (v_N^- - U_N^-, e_N) \\ &= (u_N - v_N^-, e_N) + A(v - U, \Phi) \\ &= (u_N - v_N^-, e_N) + A(v - u, \Phi) \end{aligned} \quad (57)$$

Taking $v \in V$ to be a suitable interpolant of u here, we thus obtain a representation of the error e_N in terms of an interpolation error $u - v$ and the discrete solution Φ of the associated dual problem, combined through the bilinear form $A(\cdot, \cdot)$. To obtain the a priori error estimates, we estimate below the interpolation error $u - v$ in $L_\infty(L_2(\Omega))$ and the time derivative $\dot{\Phi}$ in $L_1(L_2(\Omega))$ using discrete strong stability.

15.2 Strong stability of the discrete dual problem

We apply Lemma 3 to the function $v(t) = \Phi(T - t)$, to obtain the strong stability estimate

$$\begin{aligned} \|\Phi\|_I + \sum_{n=1}^N \int_{I_n} \{\|\dot{\Phi}\| + \|\Delta_n \Phi\|\} dt \\ + \sum_{n=1}^N \|\Phi\|_{I_n} \leq C L_N \|e_N\| \end{aligned} \quad (58)$$

with $L_N = (\log(T_N/k_n) + 1)^{1/2}$.

15.3 Proof of the a priori error estimate (36)

In the error representation, we take the interpolant to be $v \equiv \tilde{u} \equiv Q_n \pi_n u$ on I_n , where Q_n is the $L_2(I_n)$ -projection

onto polynomials of degree q on I_n and π_n is the elliptic projection defined in Section 14. For $q = 0$, we thus take

$$\tilde{u}|_{I_n} = k_n^{-1} \int_{I_n} \pi_n u \, ds \quad (59)$$

and for $q = 1$, we take

$$\begin{aligned} \tilde{u}|_{I_n} = & k_n^{-1} \int_{I_n} \pi_n u \, ds + \frac{12(t - t_{n-1} - k_n/2)}{k_n^3} \\ & \times \int_{I_n} \left(\frac{s - t_{n-1} - k_n}{2} \right) \pi_n u \, ds \end{aligned} \quad (60)$$

With this choice of interpolant, (57) reduces to

$$\begin{aligned} \|e_N\|^2 = & (u_N - \tilde{u}_N, e_N) + \sum_{n=1}^N (u - \pi_n u, \phi)_n \\ & + \sum_{n=1}^{N-1} (u_n - \tilde{u}_n, [\Phi]_n) - (u_N - \tilde{u}_N, \Phi_N^-) \end{aligned} \quad (61)$$

where we have used (50), (56) and the fact that $(\pi_n u - \tilde{u}, v)_n = 0$ for all $v \in V_n$, and thus, in particular, for $v = \phi$ and $v = \Delta_n \Phi$.

Using Lemma 5 and the fact that Q_n is bounded in $\|\cdot\|_{L_n}$, we have

$$\begin{aligned} \|u - \tilde{u}\|_{L_n} \leq & \|u - Q_n u\|_{L_n} + \|Q_n(u - \pi_n u)\|_{L_n} \\ \leq & C \left(\min_{1 \leq q \leq 1} k_n^q \|u_t^{(q)}\|_{L_n} + \|h_n^2 D^2 u\|_{L_n} \right) \end{aligned} \quad (62)$$

where the bound for $u - Q_n u$ follows from the Taylor expansion

$$\begin{aligned} u(t) = & u(t_n) + \int_{t_n}^t \dot{u}(s) \, ds \\ = & u(t_n) + (t - t_n) \dot{u}(t_n) + \int_{t_n}^t (t - s) \ddot{u}(s) \, ds \end{aligned}$$

noting that Q_n is the identity on the polynomial part of $u(t)$.

From (61) we thus obtain,

$$\begin{aligned} \|e_N\|^2 \leq & C \max_{1 \leq n \leq N} \left(\min_{1 \leq q \leq 1} k_n^q \|u_t^{(q)}\|_{L_n} + \|h_n^2 D^2 u\|_{L_n} \right) \\ & \times \left(\|e_N\| + \sum_{n=1}^N \int_{I_n} \|\phi\| \, dt + \sum_{n=1}^{N-1} \|[\Phi]_n\| + \|\Phi_N^-\| \right) \end{aligned}$$

and conclude in view of (58) that

$$\|e_N\| \leq CL_N \max_{1 \leq n \leq N} \left(\min_{1 \leq q \leq 1} k_n^q \|u_t^{(q)}\|_{L_n} + \|h_n^2 D^2 u\|_{L_n} \right) \quad (63)$$

By a local analysis this estimate extends to $\|e\|_{L_N}$, completing the proof of (36).

15.4 Proof of the a priori error estimate (37)

In the error representation formula (57), we now choose $v = R_n \pi_n u$, where R_n is the (Radau) projection onto linear functions on I_n , defined by $(R_n \pi_n u)_n^- = \pi_n u_n$ and the condition that $R_n \pi_n u - \pi_n u$ has mean value zero over I_n , that is, we take

$$R_n \pi_n u|_{I_n} = \pi_n u_n + (t - t_n) \frac{2}{k_n^2} \int_{I_n} \pi_n (u_n - u) \, ds \quad (64)$$

With this choice of interpolant, (57) reduces to

$$\begin{aligned} \|e_N\|^2 = & (u_N - \pi_N u_N, e_N) + \sum_{n=1}^N (u - \pi_n u, \phi)_n \\ & - \sum_{n=1}^N (\nabla(\pi_n u - R_n \pi_n u), \nabla \Phi)_n \\ & + \sum_{n=1}^{N-1} (u_n - \pi_n u_n, [\Phi]_n) - (u_N - \pi_N u_N, \Phi_N^-) \end{aligned} \quad (65)$$

where in the first sum we have used the fact that $\pi_n u - R_n \pi_n u$ is orthogonal to Φ (which is constant in t on I_n), and in the second sum we have used (50). For the latter term, we have

$$\begin{aligned} (\nabla(\pi_n u - R_n \pi_n u), \nabla \Phi)_n = & (\nabla(\pi_n u - R_n \pi_n u), \nabla \Phi_n^-)_n \\ & + (\nabla(\pi_n u - R_n \pi_n u), (t - t_n) \nabla \Phi)_n \end{aligned}$$

so that by our choice of $R_n \pi_n u$,

$$\begin{aligned} |(\nabla(\pi_n u - R_n \pi_n u), \nabla \Phi)_n| &= |(\nabla(\pi_n u - R_n \pi_n u), (t - t_n) \nabla \Phi)_n| \\ &\leq k_n \|\Delta_n(\pi_n u - R_n \pi_n u)\|_{L_n} \int_{I_n} |\phi| \, dt \end{aligned} \quad (66)$$

Using Taylor expansions, we easily find that

$$\|\Delta_n(\pi_n u - R_n \pi_n u)\|_{L_n} \leq C k_n^2 \|\Delta_n \pi_n u_t^{(2)}\|_{L_n} \quad (67)$$

Finally, we note that for any $w \in H^2(\Omega) \cap H_0^1(\Omega)$ we have

$$\begin{aligned} (-\Delta_n \pi_n w, v) = & (\nabla \pi_n w, \nabla v) = (\nabla w, \nabla v) = (-\Delta w, v) \\ & \forall v \in S_n \end{aligned}$$

from which we deduce by taking $v = -\Delta_n \pi_n w$ that

$$\|\Delta_n \pi_n w\| \leq \|\Delta w\| \quad \forall w \in H^2(\Omega) \cap H_0^1(\Omega) \quad (68)$$

It now follows from (65) through (68), together with Lemma 5 and strong stability for Φ , that

$$\begin{aligned} \|e_N\|^2 \leq & C \max_{1 \leq n \leq N} \left(\min_{1 \leq q \leq 2} k_n^q \|u_t^{(q)}\|_{L_n} + \|h_n^2 D^2 u\|_{L_n} \right) \\ & \times \left(\|e_N\| + \sum_{n=1}^N \int_{I_n} \|\phi\| \, dt + \sum_{n=1}^{N-1} \|[\Phi]_n\| + \|\Phi_N^-\| \right) \\ \leq & \|e_N\| CL_N \max_{1 \leq n \leq N} \left(\min_{1 \leq q \leq 2} k_n^q \|u_t^{(q)}\|_{L_n} + \|h_n^2 D^2 u\|_{L_n} \right) \end{aligned}$$

where we have used the notation $u_t^{(3)} = \Delta \tilde{u}$. This completes the proof of (37) and Theorem 1.

16 PROOF OF THE A POSTERIORI ERROR ESTIMATES

The proof of the a posteriori error estimates is similar to that of the a priori error estimates just presented. The difference is that now the error representation involves the exact solution ϕ of the continuous dual problem (55), together with the residual of the discrete solution U . By the definition of the dual problem and the bilinear form A , we have

$$\|e_N\|^2 = A(e, \phi) = A(u, \phi) - A(U, \phi)$$

Using now that

$$A(u, \phi) = (u^0, \phi_0^+) + (f, \phi)_I$$

together with the Galerkin orthogonality, we obtain the following error representation in terms of the discrete solution U , the dual solution ϕ , and data u^0 and f :

$$\begin{aligned} \|e_N\|^2 = & A(U, v - \phi) + (u^0, \phi_0^+ - v_0^+) + (f, \phi - v)_I \\ = & \sum_{n=1}^N ((U, v - \phi)_n + (\nabla U, \nabla(v - \phi))_n) \\ & + \sum_{n=1}^N ((U)_{n-1}, v_{n-1}^+ - \phi_{n-1}^+) + (f, \phi - v)_I \\ = & I + II + III \end{aligned} \quad (69)$$

with obvious notation. This holds for all $v \in V$.

To prove (38), we now choose $v|_{I_n} = \tilde{\phi} = Q_n P_n \phi$ in (69), and note that

$$(\dot{U}, \tilde{\phi} - \phi)_n = 0$$

Since also

$$\begin{aligned} (\nabla U, \nabla(\tilde{\phi} - P_n \phi))_n = & (-\Delta_n U, \tilde{\phi} - P_n \phi)_n \\ = & (-\Delta_n U, (Q_n - I) P_n \phi) \\ = & 0 \end{aligned}$$

it follows that

$$\begin{aligned} I = & \sum_{n=1}^N (\nabla U, \nabla(P_n - I)\phi)_n \\ = & \sum_{n=1}^N (\nabla U, \nabla(P_n - I) \int_{I_n} \phi \, dt) \end{aligned}$$

Using that

$$\Delta \int_{I_n} \phi \, dt = \int_{I_n} \Delta \phi \, dt = \int_{I_n} \phi(t_n) - \phi(t_{n-1})$$

together with Lemma 4 and elliptic regularity, we get

$$\begin{aligned} |I| \leq & C \sum_{n=1}^N \|h_n^2 D_n^2 U\|_{L_n} \|\Delta \int_{I_n} \phi \, dt\| \\ \leq & C \max_{1 \leq n \leq N} \|h_n^2 D_n^2(U)\|_{L_n} \left(\int_0^{t_{N-1}} \|\dot{\phi}\| \, dt + 2\|\phi\|_{t_N} \right) \end{aligned} \quad (70)$$

To estimate II , we note that by Lemma 4 we have

$$|(U)_{n-1}, (P_n - I)\phi_{n-1}^+| \leq C \|h_n^2 (U)_{n-1}\| \|\Delta \phi_{n-1}^+\| \quad (71)$$

noting that the left-hand side is zero if $S_{n-1} \subset S_n$. By obvious stability and approximation properties of the $L_2(I_n)$ -projections onto the set of constant functions on I_n , we also have

$$\|\tilde{\phi} - P_n \phi\|_{L_n} \leq \|P_n \phi\|_{L_n} \leq \|\phi\|_{L_n} \quad (72)$$

and

$$\|\tilde{\phi} - P_n \phi\|_{L_n} \leq \int_{I_n} \|P_n \phi\| \, dt \leq \int_{I_n} \|\phi\| \, dt \quad (73)$$

We thus conclude that

$$|II| \leq C \max_{1 \leq n \leq N} \left\| \frac{h_n^2 (U)_{n-1}}{k_n} \right\| \sum_{n=1}^N k_n \|\Delta \phi_{n-1}^+\|$$

$$+ \max_{1 \leq n \leq N} \|U\|_{n-1} \left(\int_0^{t_{n-1}} \|\dot{\phi}\| dt + \|\phi\|_{t_n} \right) \quad (74)$$

The data term III is estimated similarly. We finally use strong stability of ϕ in the form

$$\begin{aligned} \sum_{n=1}^{N-1} k_n \|w_{n-1}^+\| &\leq \int_0^{t_{N-1}} \|w\| dt \\ &\leq \left(\int_0^{t_{N-1}} (t_N - t)^{-1} dt \right)^{1/2} \\ &\quad \times \left(\int_0^{t_N} (t_N - t) \|w\|^2 dt \right)^{1/2} \\ &\leq \frac{1}{2} \left(\log \frac{t_N}{k_N} \right)^{1/2} \|e_N\| \quad (75) \end{aligned}$$

for $w = \dot{\phi}$ and $w = \Delta\phi$, together with the estimate $k_N \|\Delta\phi_{N-1}^+\| \leq \exp(-1) \|e_N\|$.

Combining the estimates completes the proof of the posteriori error estimate (38). The proof of the a posteriori error estimate (39) is similar.

17 EXTENSION TO SYSTEMS OF CONVECTION-DIFFUSION-REACTION PROBLEMS

In a natural way, we may directly extend the scope of methods and analysis to systems of convection-diffusion-reaction equations of the form

$$\begin{cases} \dot{u} - \nabla \cdot (a \nabla u) + (b \cdot \nabla) u - f(u) = 0 & \text{in } \Omega \times (0, T] \\ \partial_n u = 0 & \text{on } \Gamma \times (0, T] \\ u(\cdot, 0) = u^0 & \text{in } \Omega \end{cases} \quad (76)$$

where $u = (u_1, \dots, u_d)$ is a vector of concentrations, $a = a(x, t)$ is a diagonal matrix of diffusion coefficients, $b = b(x, t)$ is a given convection velocity, $f(u)$ models reactions, and $\partial_n = n \cdot \nabla$ with n the exterior normal of Γ . Depending on the size of the coefficients a and b and the reaction term, this problem may exhibit more or less parabolic behavior determined by the size of the strong stability factor coupled to the associated linearized dual problem (here linearized at the exact solution u):

$$\begin{cases} -\phi - \nabla \cdot (a \nabla \phi) - \nabla \cdot (\phi b) & \text{in } \Omega \times [0, T] \\ -(f'(u))^T \phi = 0 & \\ \partial_n \phi = 0 & \text{on } \Gamma \times [0, T] \\ \phi(\cdot, T) = \psi & \text{in } \Omega \end{cases} \quad (77)$$

where $(\nabla \cdot (\phi b))_i = \nabla \cdot (\phi b_i)$ for $i = 1, \dots, d$.

18 EXAMPLES OF REACTION-DIFFUSION PROBLEMS

We now present solutions to a selection of reaction-diffusion problems, including solutions of the dual backward problem and computation of stability factors.

18.1 Moving heat source

In Figures 10 and 11, we display mesh and solution at two different times for the adaptive cG(1)dG(0) method applied to the heat equation with a moving heat source producing a moving hot spot. We notice that the space mesh adapts to the solution.

18.2 Adaptive time steps for the heat equation

We consider again the heat equation $\dot{u} - \Delta u = f$ with homogeneous Dirichlet boundary conditions on the unit square $(0, 1) \times (0, 1)$ over the time interval $[0, 100]$. The source $f(x, t) = 2\pi^2 \sin(\pi x_1) \sin(\pi x_2) [\sin(\pi^2 t) + \cos(2\pi^2 t)]$ is periodic in time, with corresponding exact solution

$$u(x, t) = \sin(\pi x_1) \sin(\pi x_2) \sin(2\pi^2 t)$$

In Figure 12, we show a computed solution using the cG(1)dG(0) method, and we also plot the time evolution of the L_2 -error in space, together with the sequence

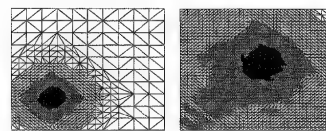


Figure 10. Meshes for moving source problem.

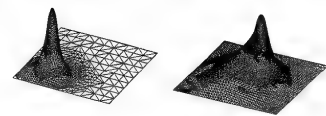


Figure 11. Solution for moving source problem.

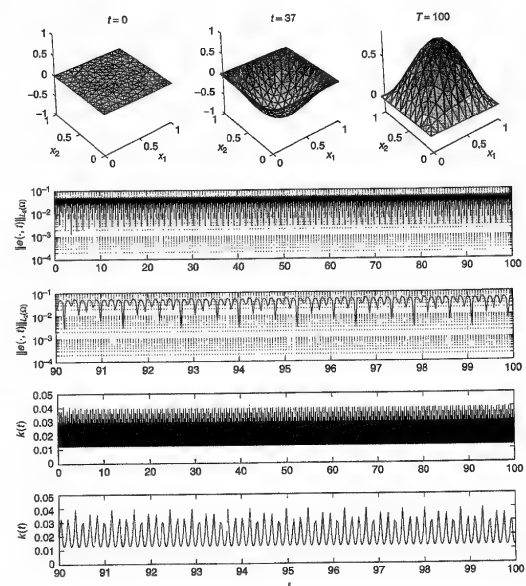


Figure 12. Heat equation: solution, error, and adaptive time steps. A color version of this image is available at <http://www.mrw.interscience.wiley.com/cecm>

of adaptive time steps. We notice that the error does not grow with time, reflecting the parabolic nature of the problem. We also note the periodic time variation of the time steps, reflecting the periodicity of the solution, with larger time steps when the solution amplitude is small.

$$\begin{cases} \dot{u} - \epsilon \Delta u = u(1-u) & \text{in } \Omega \times (0, T] \\ \partial_n u = 0 & \text{on } \Gamma \times (0, T] \\ u(\cdot, 0) = u^0 & \text{in } \Omega \end{cases} \quad (78)$$

with $\Omega = (0, 1) \times (0, 1)$, $T = 10$, $\epsilon = 0.01$, and

$$u^0(x) = \begin{cases} 0, & 0 < x_1 < 0.5 \\ 1, & 0.5 \leq x_1 < 1 \end{cases} \quad (79)$$

18.3 Logistics reaction-diffusion

We now consider the heat equation with a nonlinear reaction term, referred to as the *logistics problem*:

Through the combined action of the diffusion and reaction the solution $u(x, t)$ tends to 1 for all x with increasing time: see Figure 13. We focus interest at final time T to a circle of radius $r = 0.25$ centered at $x = (0.5, 0.5)$. The

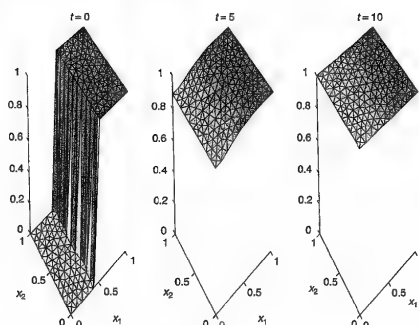


Figure 13. The logistics problem: solution at three different times. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

corresponding dual problem linearized at the exact solution u is given by

$$\begin{cases} -\phi - \epsilon \Delta \phi = (1 - 2u)\phi & \text{in } \Omega \times [0, T] \\ \partial_\nu \phi = 0 & \text{on } \Gamma \times [0, T] \\ \phi(\cdot, T) = \psi & \text{in } \Omega \end{cases} \quad (80)$$

where we choose $\psi = 1/\pi r^2$ within the circle and zero outside. In Figure 14, we plot the dual solution $\phi(\cdot, t)$ and also the stability factor $S_c(T, \psi)$ as function of T . As in the Akzo–Nobel problem discussed above, we note that $S_c(T, \psi)$ reaches a maximum for $T \sim 1$ and then decays somewhat for larger T . The decay with larger T can be understood from the sign $(1 - 2u)$ of the coefficient of the ϕ -term in the dual problem, which is positive when $u(x, t) < 0.5$, and thus is positive for small t and $x_1 < 0.5$ and negative for larger t . The growth phase in $\psi(\cdot, t)$ thus occurs after a longer phase of decay if T is large, and thus $S_c(T, \psi)$ may effectively be smaller for larger T , although the interval of integration is longer for large T .

18.4 Moving reaction front

Next, we consider a system of reaction–diffusion equations, modeling an auto-catalytic reaction, where A reacts to form B with B as a catalyst:



With u_1 the concentration of A and u_2 that of B , the system takes the form

$$\begin{cases} \dot{u}_1 - \epsilon \Delta u_1 = -u_1 u_2^2 \\ \dot{u}_2 - \epsilon \Delta u_2 = u_1 u_2^2 \end{cases} \quad (82)$$

on $\Omega \times (0, 100]$ with $\Omega = (0, 1) \times (0, 0.25)$, $\epsilon = 0.0001$ and homogeneous Neumann boundary conditions. As initial conditions, we take

$$u_1(x, 0) = \begin{cases} 0, & 0 < x_1 < 0.25 \\ 1, & 0.25 \leq x_1 < 1 \end{cases} \quad (83)$$

and $u_2(\cdot, 0) = 1 - u_1(\cdot, 0)$. The solution $u(x, t)$ corresponds to a reaction front, starting at $x_1 = 0.25$ and propagating to the right in the domain until all of A is consumed and the concentration of B is $u_2 = 1$ in all of Ω ; see Figure 15.

The dual problem, linearized at $u = (u_1, u_2)$, is given by

$$\begin{cases} -\phi_1 - \epsilon \Delta \phi_1 = -u_2^2 \phi_1 + u_2^3 \phi_2 \\ -\phi_2 - \epsilon \Delta \phi_2 = -2u_1 u_2 \phi_1 + 2u_1 u_2 \phi_2 \end{cases} \quad (84)$$

As in the previous example, we take the final time data ψ_1 for the first component of the dual to be an approximation of a Dirac delta function centered in the middle of the domain, and $\psi_2 = 0$.

We note that the stability factor peaks at the time of active reaction and that before and after the reaction front has swept the region of observation the stability factor $S_c(T, \psi)$ is significantly smaller (see Figure 16).

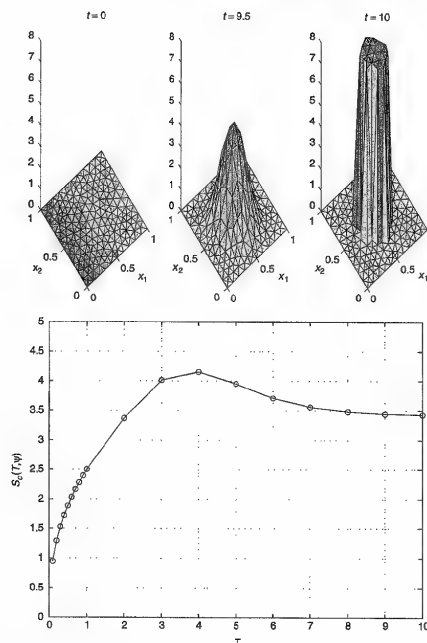


Figure 14. The logistics problem: dual solution and stability factor $S_c(T, \psi)$. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

19 COMPARISON WITH THE STANDARD APPROACH TO TIME STEP CONTROL FOR ODEs

We now compare our methods for adaptive error control with the methods for automatic time step control developed within the ODE community as presented, for example, in Hairer and Wanner (1996) and Deufelhard and Bornemann

(2002), which we refer to as the *standard approach*. The corner-stone of the standard approach is a concept of *local error*, which is (an estimate of) the error contribution from each time step, and the time step is then controlled so as to keep the local error within a certain *local error tolerance* tol. The relation between the local tolerance tol and the global error after many time steps is then left to the (intelligent) user to figure out.

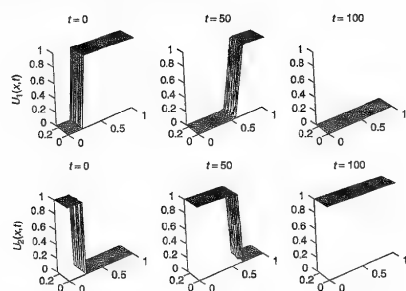


Figure 15. Reaction front problem: solution for the two components at three different times. A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

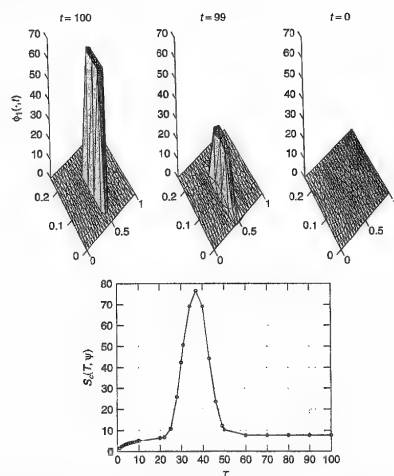


Figure 16. Reaction front problem: dual solution and stability factor $S_e(T)$ as function of T . A color version of this image is available at <http://www.mrw.interscience.wiley.com/ecn>

For the comparison, we focus on the backward Euler method and first recall our basic a posteriori error estimate (7) for the global error $e(T)$, that is,

$$\|e(T)\| \leq S_e(T) \max_{1 \leq n \leq N} \|U(t_n) - U(t_{n-1})\| \quad (85)$$

where $S_e(T)$ is the strong stability factor. We thus obtain control of the global error $\|e(T)\| \leq \text{TOL}$ if we control the local time step so that

$$\|U(t_n) - U(t_{n-1})\| \sim \text{TOL}$$

assuming $S_e(T) \sim 1$.

Next, the local error \hat{e}_n at time step n in the standard approach is defined as the difference between the first-order backward Euler solution $U(t_n)$ and an approximate solution obtained after one time step starting from $U(t_{n-1})$ and using a higher-order method. With a trapezoidal method as the higher-order method, we may obtain

$$\hat{e}_n = \frac{k_n}{2} \|f(U(t_n)) - f(U(t_{n-1}))\|$$

The standard time step control would then consist in choosing the time step so that

$$\hat{e}_n \approx \text{tol}$$

where tol would thus be the local error tolerance. To estimate the global error $e(T)$ following the standard approach, it would appear that one would have to sum all the local errors to get

$$\|e(T)\| \leq \sum_{n=1}^N \hat{e}_n \quad (86)$$

It would then seem that we should choose tol so that $\text{TOL} = N \text{tol}$, where N is the number of time steps. But as we said, the relation between the local and global error tolerance is left to the user in the standard approach.

To bring out the essential difference between our estimate of the global error (85) with a standard estimate (86), we compare the following a priori analog of the standard estimate (86), obtained using that $(d/dr)f(u) = \tilde{u}$ (and discarding a factor $1/2$):

$$\|e(T)\| \leq E_1 = \int_0^T k(t) \|\tilde{u}(t)\| dt \quad (87)$$

and the analog of our estimate (85):

$$\|e(T)\| \leq E_2 = \max_{0 \leq t \leq T} \|k(t) \tilde{u}(t)\| \quad (88)$$

where $k(t) = k_n$ on I_n , and we put $S_e(T) = 1$. Now it is clear that by integration $E_2 \leq E_1$ (assuming for simplicity that $\tilde{u}(0) = 0$) and neglecting $k\tilde{u}$, while we may have $E_2 \ll E_1$ in particular if T is large. This shows that our estimate may be very much more economical to use than the standard estimate. We note that the standard estimate will look the same independent of the function $f(u)$, and thus does not take any particular advantage of the special properties of stiff or parabolic problems, as our estimate does. This explains the possibly very suboptimal nature of the standard estimate. Notice again that our estimate does not depend on heavy exponential decay of all components: in the model problem $\tilde{u} + A\tilde{u} = 0$, we assume that A is positive semidefinite and not strictly positive definite.

In the standard approach, it is further common to choose the values of the higher-order method rather than the lower-order method because of obvious reasons. With this perspective, the estimate of the local error would then rather come out as the difference after one time step of a higher-order method, which is used in the computation and a lower-order method, which is used only for the local error estimation. Doing so we would seem to lack an error estimate for the higher-order method since after all we estimated the local error of the lower-order method.

Nevertheless, using the values of the higher-order method, it seems (surprisingly so) that in many problems including stiff problems, the global error tolerance TOL would be of the same size as the local error tolerance tol , as if there were no error accumulation. By using the results of the higher-order method, it thus seems to be possible to get global error control on the level of the local error tolerance. But no explanation of this 'miracle' has been given for the standard approach.

Well, let's see if we can explain this mystery using our sharp error analysis. Let us then take as the higher-order method the first-order backward Euler method and as the lower-order method the zero order (trivial) method: $\tilde{U}(t_n) = U(t_{n-1})$. The difference between the higher-order and the lower-order method at time step n would then simply be $\|U(t_n) - U(t_{n-1})\|$, and the time step would be chosen so that the local error $\|U(t_n) - U(t_{n-1})\| \leq \text{tol}$. But this is the same as our time-step control guaranteeing the global error control $\|e(T)\| \leq \text{tol}$ (assuming $S_e(T) = 1$). We have thus proved that for parabolic or stiff problems by using the values of the higher-order method, the global error would come out on the level of the local error tolerance, independent of the length of the simulation, and the 'miracle' would thus be possible to understand.

20 SOFTWARE

There is a variety of software available for stiff ODEs such as LSODE and RADAU5 using the standard approach to error control. The methods presented in this article have been implemented in a general multiadaptive framework as a part of the DOLFIN project developed by J. Hoffman and A. Logg and presented at <http://www.phichalmers.se/dolfin/>.

REFERENCES

- Delfour M, Hager W and Trochu F. Discontinuous Galerkin methods for ordinary differential equations. *Math. Comput.* 1981; **36**:453–473.
- Deufelhard P, Bornemann F and Wanner G. *Scientific Computing with Ordinary Differential Equations*, Springer Texts in Applied Mathematics 42, 2002.
- Douglas Jr. J, Dupont T and Wheeler M. Galerkin methods for parabolic equations. *SIAM J. Numer. Anal.* 1970; **7**:575–626.
- Eriksson K, Johnson C and Logg A. Explicit time-stepping for stiff ODEs. *SIAM J. Sci. Comput.* 2003; **25**:1142–1157.
- Hairer E and Wanner G. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer Series in Computational Mathematics 14, 1996.
- Logg A. Multi-adaptive Galerkin methods for ODEs I. *SIAM J. Sci. Comput.* 2003; **24**:1879–1902.
- Logg A. Multi-adaptive Galerkin methods for ODEs II: Implementation & applications. *SIAM J. Sci. Comput.* 2003; **25**:1119–1144.
- Thomée V. *Galerkin Finite Element Methods for Parabolic Problems*, Springer Series in Computational Mathematics 25, 1997.
- Eriksson K and Johnson C. Adaptive finite element methods for parabolic problems I: A linear model problem. *SIAM J. Numer. Anal.* 1991; **28**:43–77.
- Eriksson K and Johnson C. Adaptive finite element methods for parabolic problems II: Optimal error estimates in L_∞ , L_2 and $L_{\infty}L_{\infty}$. *SIAM J. Numer. Anal.* 1995; **32**:706–740.
- Eriksson K and Johnson C. Adaptive finite element methods for parabolic problems IV: Nonlinear problems. *SIAM J. Numer. Anal.* 1995; **32**:1729–1749.
- Eriksson K and Johnson C. Adaptive finite element methods for parabolic problems V: Long-time integration. *SIAM J. Numer. Anal.* 1995; **32**:1750–1763.
- Eriksson K, Estep D and Johnson C. *Computational Differential Equations*. Studentlitteratur, Lund, 1996.
- Eriksson K, Johnson C and Larsson S. Adaptive finite element methods for parabolic problems VI: Analytic semigroups. *SIAM J. Numer. Anal.* 1998; **35**:1315–1325.
- Eriksson K, Johnson C and Thomée V. Time discretization of parabolic problems by the discontinuous Galerkin method. *RAIRO Man* 1985; **19**:611–643.
- Eriksson K, Estep D, Hansbo P and Johnson C. Introduction to adaptive methods for differential equations. *Acta Numerica*. Cambridge University Press, 1995; 105–158.
- Johnson C. Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations. *SIAM J. Numer. Anal.* 1988; **25**:908–926.
- Johnson C and Logg A. *Applied Mathematics: Body and Soul V: Dynamical Systems*. To appear at Springer, 2005.

FURTHER READING

Chapter 25

Time-dependent Problems with the Boundary Integral Equation Method

Martin Costabel

IRMAR, Université de Rennes 1, Campus de Beaulieu, Rennes, France

| | |
|---------------------------------|-----|
| 1 Introduction | 703 |
| 2 Space-time Integral Equations | 705 |
| 3 Laplace Transform Methods | 713 |
| 4 Time-stepping Methods | 714 |
| References | 719 |

1 INTRODUCTION

Like stationary or time-harmonic problems, transient problems can be solved by the boundary integral equation method. When the material coefficients are constant, a fundamental solution is known and the data are given on the boundary, the reduction to the boundary provides efficient numerical methods, in particular for problems posed on unbounded domains.

Such methods are widely and successfully being used for numerically modeling problems in heat conduction and diffusion, in the propagation and scattering of acoustic, electromagnetic, and elastic waves, and in fluid dynamics.

One can distinguish three approaches to the application of boundary integral methods on parabolic and hyperbolic initial-boundary value problems: space-time integral equations, Laplace-transform methods, and time-stepping methods.

1. *Space-time integral equations* use the fundamental solution of the parabolic or hyperbolic partial differential equations.

The construction of the boundary integral equations via representation formulas and jump relations, the appearance of single- and double-layer potentials, and the classification into first- and second-kind integral equations follow in a large part the formalism known for elliptic problems. Causality implies that the integral equations are of Volterra type in the time variable, and time-invariance implies that they are of convolution type in time.

Numerical methods constructed from these space-time boundary integral equations are global in time, that is, they compute the solution in one step for the entire time interval. The boundary is the lateral boundary of the space-time cylinder and therefore has one dimension more than the boundary of the spatial domain. This increase in dimension at first means a substantial increase in complexity:

- To compute the solution for a certain time, one needs the solution for all the preceding times since the initial time.
- The system matrix is much larger.
- The integrals are higher-dimensional. For a problem with three space dimensions, the matrix elements in a Galerkin method can require six-dimensional integrals.

While the increase in memory requirements for the storage of the solution for preceding times cannot completely be avoided, there are situations in which the other two reasons for increased complexity are, in part, neutralized by special features of the problem:

- The system matrix has a special structure related to the Volterra structure (finite convolution in time) of the integral equations. When low-order basis functions in time are used, the matrix is of block-triangular Toeplitz form, and for its inversion, only one block – which has the size of the system matrix for a corresponding time-independent problem – needs to be inverted.
- When a strong Huygens principle is valid for the partial differential equation, the integration in the integral representation is not extended over the whole lateral boundary of the space-time cylinder, but only over its intersection with the surface of the backward light cone. This means, firstly, that the integrals are of the same dimensionality as for time-independent problems, and secondly, that the dependence is not extended arbitrarily far into the past, but only up to a time corresponding to the time of traversal of the boundary with the fixed finite propagation speed. These 'retarded potential integral equations' are of importance for the scalar wave equation in three-space dimensions, and, to a certain extent, for equations derived from them in electromagnetics and elastodynamics. On the other hand, such a Huygens principle is not valid for the wave equation in a two-space dimension, or for the heat equation or for problems in elastodynamics or in fluid dynamics.

2. *Laplace transform methods* solve frequency-domain problems, possibly for complex frequencies. For each fixed frequency, a standard boundary integral method for an elliptic problem is applied, and then the transformation back to the time domain employs special methods for the inversion of Fourier or Laplace transforms. The choice of a numerical method for the inverse Laplace transform can be guided by the choice of an approximation of the exponential function corresponding to a linear multistep method for ordinary differential equations. This idea is related to the *operational quadrature method* (Lubich, 1994).

Laplace or Fourier transform is also used the other way round to pass from the time domain to the frequency domain. This can be done using fast Fourier transform (FFT) in order to simultaneously solve problems for many frequencies from one time-domain computation, or one can solve a time-domain problem with a time-harmonic right-hand side to get the solution for one fixed frequency. It has been observed that this can be efficient too (Sayah, 1998) because of less strict requirements for the spatial resolution.

3. *Time-stepping methods* start from a time discretization of the original initial-boundary value problem via an implicit scheme and then use boundary integral equations to solve the resulting elliptic problems for each time step. Here, the difficulty lies in the form of the problem for one time step, which has nonzero initial data and thus is not in the ideal form for an application of the boundary

integral method, namely vanishing initial conditions and volume forces and nonhomogeneous boundary data. The solution after a time step, which defines the initial condition for the next time step, has no reason to vanish inside the domain. Various methods have been devised to overcome this problem:

Using volume potentials to incorporate the nonzero initial data often is not desirable since it requires discretization of the domain and thus defines the advantage of the reduction to the boundary. Instead of a volume potential (Newton potential), another particular solution (or approximate particular solution) of the stationary problem can be used. This particular solution may be obtained by fast solution methods, for example FFT or a fast Poisson solver on a fictitious domain, or by meshless discretization of the domain using special basis functions like thin-plate splines or other radial basis functions (so-called *dual reciprocity method*); see Aliabadi and Wrobel (2002).

Another idea is to consider not a single time step, but all time steps up to the final time together as a discrete convolution equation for the sequence of solutions at the discrete time values. Such a discrete convolution operator whose (time-independent) coefficients are elliptic partial differential operators has a fundamental solution, which can then be used to construct a pure boundary integral method for the solution of the time-discretized problem. A fundamental solution, which is also a discrete convolution operator, can be given explicitly for simple time-discretization schemes like the backward Euler method ('Rothe method', Chapko and Kress, 1997). For a whole class of higher-order one-step or multistep methods, it can be constructed using Laplace transforms via the *operational quadrature method* (Lubich and Schneider, 1992; Lubich, 1994).

These three approaches for the construction of boundary integral methods cannot be separated completely. There are many points of overlap:

The space-time integral equation method leads, after discretization, to a system that has the same finite time convolution structure that one gets from time-stepping schemes. The main difference is that the former needs the knowledge of a space-time fundamental solution. But this is simply the inverse Laplace transform of the fundamental solution of the corresponding time-harmonic problem.

The Laplace transform appears in several roles. It can be used to translate between the time domain and the frequency domain at the level of the formulation of the problem, and also at the level of the solution.

The stability analysis for all known algorithms, for the space-time integral equation methods as for the time-stepping methods, passes by the transformation to the frequency domain and corresponding estimates for the stability

of boundary integral equations methods for elliptic problems. The difficult part in this analysis is to find estimates uniform with respect to the frequency.

For *parabolic* problems, some analysis of integral equation methods and their numerical realization has been known for a long time, and the classical results for second-kind integral equations on smooth boundaries are summarized in the book by Pogorzelski (1966). All the standard numerical methods available for classical Fredholm integral equations of the second kind, like collocation methods or Nyström methods, can be used in this case.

More recently, variational methods have been studied in a setting of anisotropic Sobolev spaces that allow the coverage of first-kind integral equations and nonsmooth boundaries. It has been found that, unlike the parabolic partial differential operator with its time-independent energy and no regularizing property in time direction, the first-kind boundary integral operators have a kind of anisotropic space-time ellipticity (Costabel, 1990; Arnold and Noon, 1989; Brown, 1989; Brown and Shen, 1993).

This ellipticity leads to unconditionally stable and convergent *Galerkin methods* (Costabel, 1990; Arnold and Noon, 1989; Hsiao and Saranen, 1993; Hebekker and Hsiao, 1993). Because of their simplicity, *collocation methods* are frequently used in practice for the discretization of space-time boundary integral equations. An analysis of collocation methods for second-kind boundary integral equations for the heat equation was given by Costabel *et al.* (1987). Fourier analysis techniques for the analysis of stability and convergence of collocation methods for parabolic boundary integral equations, including first-kind integral equations, have been studied more recently by Hamina and Saranen (1994) and by Costabel and Saranen (2000, 2001, 2003).

The operational quadrature method for parabolic problems was introduced and analyzed by Lubich and Schneider (1992).

For *hyperbolic* problems, the mathematical analysis is mainly based on variational methods as well (Bamberger and Ha Duong, 1986; Ha-Duong, 1990, 1996). There is now a lack of ellipticity, which on the one hand leads to a loss of an order of regularity in the error estimates. On the other hand, most coercivity estimates are based on a passage to complex frequencies, which may lead to stability constants that grow exponentially in time. Instabilities (that are probably unrelated to this exponential growth) have been observed, but their analysis does not seem to be complete (Becache, 1991; Peirce and Siebrits, 1996; Peirce and Siebrits, 1997; Birgisson *et al.*, 1999). Analysis of variational methods exists for the main domains of application of space-time boundary integral equations, first of all for the scalar wave equation,

where the boundary integrals are given by retarded potentials, and also for elastodynamics (Becache, 1993; Becache and Ha-Duong, 1994; Chudinovich, 1993c; Chudinovich, 1993b; Chudinovich, 1993a), piezoelectricity (Khutoryan-sky and Sosa, 1995), and for electrodynamics (Bachelot and Lange, 1995; Bachelot *et al.*, 2001; Rynne, 1999; Chudinovich, 1997). An extensive review of results on variational methods for the retarded potential integral equations is given by Ha-Duong (2003).

As in the parabolic case, collocation methods are practically important for the hyperbolic space-time integral equations. For the retarded potential integral equation, the stability and convergence of collocation methods has now been established (Davies, 1994, 1998; Davies and Duncan, 1997, 2003).

Finally, let us mention that there have also been important developments in the field of *fast methods* for space-time boundary integral equations (Michielssen, 1998; Jiao *et al.*, 2002; Michielssen *et al.*, 2000; Greengard and Strain, 1990; Greengard and Lin, 2000).

2 SPACE-TIME INTEGRAL EQUATIONS

2.1 Notations

We will now study some of the above-mentioned ideas in closer detail. Let $\Omega \subset \mathbb{R}^n$, ($n \geq 2$), be a domain with compact boundary Γ . The outer normal vector is denoted by \mathbf{n} and the outer normal derivative by ∂_n .

Let $T > 0$ be fixed. We denote by Q the space-time cylinder over Ω and Σ its lateral boundary:

$$\begin{aligned} Q &= (0, T) \times \Omega \\ \Sigma &= (0, T) \times \Gamma \\ \partial Q &= (\{0\} \times \bar{\Omega}) \cup \Sigma \cup (\{T\} \times \bar{\Omega}) \end{aligned}$$

For the description of the general principles, we consider only the simplest model problem of each type. We also assume that the right-hand sides have the right structure for the application of a 'pure' boundary integral method. The volume sources and the initial conditions vanish, so that the whole system is driven by boundary sources.

Elliptic problem (Helmholtz equation with frequency $\omega \in \mathbb{C}$):

$$\begin{aligned} (\Delta + \omega^2)u &= 0 \quad \text{in } \Omega \\ u &= g \quad (\text{Dirichlet}) \quad \text{or } \partial_n u = h \quad (\text{Neumann}) \quad \text{on } \Gamma \quad (\mathcal{C}) \\ &\text{radiation condition at } \infty \end{aligned}$$

Parabolic problem (heat equation):

$$\begin{aligned} (\partial_t - \Delta)u &= 0 \quad \text{in } Q \\ u &= g \text{ (Dirichlet)} \quad \text{or } \partial_n u = h \text{ (Neumann)} \quad \text{on } \Sigma \quad (\mathcal{P}) \\ u &= 0 \quad \text{for } t \leq 0 \end{aligned}$$

Hyperbolic problem (wave equation with velocity $c > 0$):

$$\begin{aligned} (c^{-2}\partial_t^2 - \Delta)u &= 0 \quad \text{in } Q \\ u &= g \text{ (Dirichlet)} \quad \text{or } \partial_n u = h \text{ (Neumann)} \quad \text{on } \Sigma \quad (\mathcal{H}) \\ u &= 0 \quad \text{for } t \leq 0 \end{aligned}$$

2.2 Space-time representation formulas

2.2.1 Representation formulas and jump relations

The derivation of boundary integral equations follows from a general method that is valid (under suitable smoothness hypotheses on the data) in the same way for all three types of problems. In fact, what counts for (\mathcal{P}) and (\mathcal{H}) is the property that the lateral boundary Σ is noncharacteristic.

The first ingredient for a boundary element method (BEM) is a fundamental solution G . As an example, in three dimensions, we have, respectively:

$$G_\omega(x) = \frac{e^{i\omega|x|}}{4\pi|x|} \quad (\mathcal{E})$$

$$G(t, x) = \begin{cases} (4\pi t)^{-3/2} e^{-|x|^2/(4t)} & (t \geq 0) \\ 0 & (t \leq 0) \end{cases} \quad (\mathcal{P})$$

$$G(t, x) = \frac{1}{4\pi|x|} \delta\left(t - \frac{|x|}{c}\right) \quad (\mathcal{H})$$

Representation formulas for a solution u of the homogeneous partial differential equation and $x \notin \Gamma$ are obtained from Green's formula, applied with respect to the space variables in the interior and exterior domain. We assume that u is smooth in the interior and the exterior up to the boundary, but has a jump across the boundary. The jump of a function v across Γ is denoted by $[v]$:

$$u(x) = \int_\Gamma \{\partial_{n(y)} G(x-y) u(y)\} - G(x-y) [\partial_n u(y)] d\sigma(y) \quad (\mathcal{E})$$

$$u(t, x) = \int_0^t \int_\Gamma \{\partial_{n(y)} G(t-s, x-y) u(s, y)\} - G(t-s, x-y) [\partial_s u(s, y)] d\sigma(y) ds \quad (\mathcal{P})$$

$$\begin{aligned} u(t, x) &= \int_0^t \int_\Gamma \{\partial_{n(y)} G(t-s, x-y) u(s, y)\} \\ &\quad - G(t-s, x-y) [\partial_s u(s, y)] d\sigma(y) ds \\ &= \int_\Gamma \left\{ \partial_{n(y)} \frac{1}{4\pi|x-y|} \left[u\left(t - \frac{|x-y|}{c}, y\right) \right] \right. \\ &\quad \left. - \frac{\partial_{n(y)}|x-y|}{4\pi c|x-y|} \left[\partial_s u\left(t - \frac{|x-y|}{c}, y\right) \right] \right. \\ &\quad \left. - \frac{1}{4\pi|x-y|} \left[\partial_s u\left(t - \frac{|x-y|}{c}, y\right) \right] d\sigma(y) \right\} \quad (7c) \end{aligned}$$

Thus, the representation in the parabolic case uses integration over the past portion of Σ in the form of a finite convolution over the interval $[0, t]$, whereas in the hyperbolic case, only the intersection of the interior of the backward light cone with Σ is involved. In 3D, where Huyghens' principle is valid for the wave equation, the integration extends only over the boundary of the backward light cone, and the last formula shows that the integration can be restricted to Γ , giving a very simple representation by 'retarded potentials'.

We note that in the representation by retarded potentials, all those space-time points (s, y) contribute to $u(t, x)$ from where the point (t, x) is reached with speed c by travelling through the space \mathbb{R}^3 . In the case of waves propagating in the exterior of an obstacle, this leads to the seemingly paradoxical situation that a perturbation at (s, y) can contribute to $u(t, x)$, although no signal from y has yet arrived in x , because in physical space, it has to travel around the obstacle.

All three representation formulas can be written in a unified way by introducing the single-layer potential \mathcal{S} and the double-layer potential \mathcal{D} :

$$u = \mathcal{D}([u]) - \mathcal{S}([\partial_n u]) \quad (1)$$

In all the cases, there hold the classical jump relations in the form

$$\begin{aligned} [\mathcal{D}v] &= v; & [\partial_n \mathcal{D}v] &= 0 \\ [\mathcal{S}\phi] &= 0; & [\partial_n \mathcal{S}\phi] &= -\phi \end{aligned}$$

It therefore appears natural to introduce the boundary operators from the sums and differences of the one-sided traces on the exterior (Γ^+) and interior (Γ^-) of Γ :

$$\begin{aligned} V &:= \mathcal{S}|_\Gamma & \text{(single-layer potential)} \\ K &:= \frac{1}{2}(\mathcal{D}|_{\Gamma^+} + \mathcal{D}|_{\Gamma^-}) & \text{(double-layer potential)} \\ K' &:= \frac{1}{2}(\partial_n \mathcal{S}|_{\Gamma^+} + \partial_n \mathcal{S}|_{\Gamma^-}) & \text{(normal derivative of} \\ & & \text{single-layer potential)} \\ W &:= -\partial_n \mathcal{D}|_\Gamma & \text{(normal derivative of} \\ & & \text{double-layer potential)} \end{aligned}$$

2.2.2 Boundary integral equations

In a standard way, the jump relations, together with these definitions, lead to boundary integral equations for the Dirichlet and Neumann problems. Typically, one has a choice of at least four equations for each problem: The first two equations come from taking the traces in the representation formula (1) (*direct method*), the third one comes from a *single-layer representation*

$$u = \mathcal{S}\psi \quad \text{with unknown } \psi$$

and the fourth one from a *double-layer representation*

$$u = \mathcal{D}w \quad \text{with unknown } w$$

For the exterior Dirichlet problem ($u|_\Gamma = g$ given, $\partial_n u|_\Gamma = \psi$ unknown):

$$V\psi = \left(-\frac{1}{2} + K\right)g \quad (D1)$$

$$\left(\frac{1}{2} + K'\right)\psi = -Wg \quad (D2)$$

$$V\psi = g \quad (D3)$$

$$\left(\frac{1}{2} + K\right)w = g \quad (D4)$$

For the exterior Neumann problem ($u|_\Gamma = g = v$ unknown, $\partial_n u|_\Gamma = h$ given):

$$\left(\frac{1}{2} - K\right)v = -Vh \quad (N1)$$

$$Wv = -\left(\frac{1}{2} + K'\right)h \quad (N2)$$

$$\left(\frac{1}{2} - K'\right)\psi = -h \quad (N3)$$

$$Ww = -h \quad (N4)$$

Remember that this formal derivation is rigorously valid for all three types of problems. One notes that second- and first-kind integral equations alternate nicely. For open surfaces, however, only the first-kind integral equations exist. The reason is that a boundary value problem on an open surface fixes not only a one-sided trace but also the jump of the solution; and therefore the representation formula coincides with a single-layer potential representation for the Dirichlet problem and with a double-layer potential representation for the Neumann problem.

The same abstract form of space-time boundary integral equations (D1)–(D4) and (N1)–(N4) is obtained for more general classes of second-order initial-boundary value problems. If a space-time fundamental solution is known, then Green's formulas for the spatial part of the partial differential operator are used to get the representation formulas and jump relations. The role of the normal derivative is played by the conormal derivative.

Since for time-independent boundaries the jumps across the lateral boundary Σ involve only jumps across the spatial boundary Γ at a fixed time t , the jump relations and representation formulas for a much wider class of higher-order elliptic systems (Costabel and Dauge, 1997) could be used to obtain space-time boundary integral equations for parabolic and hyperbolic initial-boundary value problems associated with such partial differential operators. In the general case, this has yet to be studied.

2.2.3 Examples of fundamental solutions

The essential requirement for the construction of a boundary integral equation method is the availability of a fundamental solution. This can be a serious restriction on the use of the space-time integral equation method because explicitly given and sufficiently simple fundamental solutions are known for far less parabolic and hyperbolic equations than for their elliptic counterparts.

In principle, one can pass from the frequency domain to the time domain by a simple Laplace transform, and therefore the fundamental solution for the time-dependent problem always has a representation by a Laplace integral of the frequency-dependent fundamental solution of the corresponding elliptic problem. In practice, this representation can be rather complicated. An example where this higher level of complexity of the time-domain representation is visible, but possibly still acceptable, is the *dissipative wave equation* with a coefficient $\alpha > 0$ (and speed $c = 1$ for simplicity)

$$(\partial_t^2 + \alpha\partial_t - \Delta)u = 0$$

In the frequency domain, we obtain the same equation as for the wave equation with ω simply replaced by $\omega_\alpha = \sqrt{(\omega^2 + i\alpha\omega)}$. The time-harmonic fundamental solution in three dimensions is therefore simply

$$G_{\omega_\alpha}(x) = \frac{1}{4\pi|x|} e^{i(x|\sqrt{\omega^2 + i\alpha\omega})}$$

From this, we obtain by inverse Laplace transformation

$$\begin{aligned} G(t, x) &= \frac{e^{-\alpha t/2}}{4\pi|x|} \left(\delta(t - |x|) + \frac{\alpha|x|}{2\sqrt{t^2 - |x|^2}} \right. \\ &\quad \left. \times I_1\left(\frac{\alpha}{2}\sqrt{t^2 - |x|^2}\right) \theta(t - |x|) \right) \end{aligned}$$

with the Dirac distribution δ , the Heaviside function θ , and the modified Bessel function I_1 . We see that there is no strong Huyghens principle, and the integrals in the corresponding space-time integral equations will be extended

over the whole intersection of the boundary Σ with the solid backward light cone $\{(s, y) \mid t - s > |x - y|\}$.

For the case of *elastodynamics*, the corresponding space-time integral equations have not only been successfully used for a long time in practice (Mansur, 1983; Antes, 1985, 1988) but they have also been studied mathematically. Isotropic homogeneous materials are governed by the second-order hyperbolic system for the n -component vector field \mathbf{u} of the displacement

$$\rho \partial_t^2 \mathbf{u} - \operatorname{div} \boldsymbol{\sigma} = 0 \quad \text{with } \sigma_{ij} = \mu(\partial_i u_j + \partial_j u_i) + \lambda \delta_{ij} \operatorname{div} \mathbf{u}$$

Here, ρ is the density and λ and μ are the Lamé constants. The role of the normal derivative ∂_n is played by the traction operator T_n where $T_n \mathbf{u} = \boldsymbol{\sigma} \cdot \mathbf{n}$ is the normal stress. The role of the Dirichlet and Neumann boundary conditions are played by the displacement and traction boundary conditions respectively:

$$\mathbf{u} = \mathbf{g} \quad (\text{displacement}) \quad \text{or } T_n \mathbf{u} = \mathbf{h} \quad (\text{traction}) \quad \text{on } \Sigma$$

In three dimensions, the space-time fundamental solution shows the longitudinal (pressure) and transversal (shear) waves that propagate with the two velocities

$$c_p = \sqrt{\frac{\lambda + 2\mu}{\rho}} \quad \text{and} \quad c_s = \sqrt{\frac{\mu}{\rho}}$$

But there is no strict Huygens principle; the support of the fundamental solution is not contained in the union of the two conical surfaces determined by these two speeds but rather in the closure of the domain between these two surfaces. The fundamental solution is a (3×3) matrix G , whose entries are given by

$$G_{jk}(t, \mathbf{x}) = \frac{1}{4\pi\rho|\mathbf{x}|^3} \times \left\{ t^2 \left[\frac{x_j x_k}{|\mathbf{x}|^2} \delta \left(t - \frac{|\mathbf{x}|}{c_p} \right) + \left(\delta_{jk} - \frac{x_j x_k}{|\mathbf{x}|^2} \right) \delta \left(t - \frac{|\mathbf{x}|}{c_s} \right) \right] + t \left(3 \frac{x_j x_k}{|\mathbf{x}|^2} - \delta_{jk} \right) \left[\theta \left(t - \frac{|\mathbf{x}|}{c_p} \right) - \theta \left(t - \frac{|\mathbf{x}|}{c_s} \right) \right] \right\}$$

Here, δ_{jk} is the Kronecker symbol, δ is the Dirac distribution, and θ is the Heaviside function.

Detailed descriptions of the space-time boundary integral equations in elastodynamics corresponding to (D1)–(D4) and (N1)–(N4) above can be found in many places (Chudinovich, 1993a, b; Becache and Ha-Duong, 1994; Brebbia *et al.*, 1984; Antes, 1988; Aliabadi and Wrobel, 2002).

Whereas the frequency-domain fundamental solution is explicitly available for generalizations of elastodynamics such as certain models of anisotropic elasticity or thermoelasticity (Kupradze *et al.*, 1979) or viscoelasticity (Schanz, 2001b), the time-domain fundamental solution quickly becomes very complicated (for an example in two-dimensional piezoelectricity, see Wang *et al.*, 2003), or is completely unavailable.

For the case of *electrodynamics*, space-time integral equations have been used and analyzed extensively, too, in the past twelve years (Pujols, 1991; Däschle, 1992; Terrasse, 1993; Bachelot and Lange, 1995; Chudinovich, 1997). An analysis of numerical methods on the basis of variational formulations is available, and the coupling of space-time integral equation methods with domain finite element methods has also been studied (Sayali, 1998; Bachelot *et al.*, 2001).

Maxwell's equations being a first-order system, the above formalism with its distinction between Dirichlet and Neumann conditions and between single- and double-layer potentials makes less sense here. There are, however, additional symmetries that allow to give a very 'natural' form to the space-time boundary integral equations and their variational formulations. The close relationship between the Maxwell equations and the scalar wave equation in three dimensions implies the appearance of retarded potentials here, too.

The system of Maxwell's equations in a homogeneous and isotropic material with electric permittivity ϵ and magnetic permeability μ is

$$\begin{aligned} \mu \partial_t \mathbf{H} + \operatorname{curl} \mathbf{E} &= 0 \\ \epsilon \partial_t \mathbf{E} - \operatorname{curl} \mathbf{H} &= 0 \end{aligned}$$

The speed of light is $c = 1/\sqrt{\epsilon\mu}$, and the corresponding retarded potential can be abbreviated as

$$S(\mathbf{u})(t, \mathbf{x}) = \frac{1}{4\pi} \int_{\Gamma} \frac{\mathbf{u} \left(t - \frac{|\mathbf{x} - \mathbf{y}|}{c}, \mathbf{y} \right)}{|\mathbf{x} - \mathbf{y}|} d\sigma(\mathbf{y})$$

Then, an analogue of representation formula (1) can be written in the following form:

$$\begin{aligned} \mathbf{E}(t, \mathbf{x}) &= -\mu S(\partial_t \mathbf{J})(t, \mathbf{x}) \\ &+ \frac{1}{\epsilon} \operatorname{grad}_x S(\partial_t^{-1} \operatorname{div}_\Gamma \mathbf{J})(t, \mathbf{x}) - \operatorname{curl} S(\mathbf{m})(t, \mathbf{x}) \\ \mathbf{H}(t, \mathbf{x}) &= -\epsilon S(\partial_t \mathbf{m})(t, \mathbf{x}) \\ &+ \frac{1}{\mu} \operatorname{grad}_x S(\partial_t^{-1} \operatorname{div}_\Gamma \mathbf{m})(t, \mathbf{x}) + \operatorname{curl} S(\mathbf{J})(t, \mathbf{x}) \end{aligned}$$

where \mathbf{J} and \mathbf{m} are the surface currents and surface charge densities given by the jumps across Σ :

$$[\mathbf{J}] = [\mathbf{H} \wedge \mathbf{n}] ; \quad [\mathbf{m}] = [\mathbf{n} \wedge \mathbf{E}]$$

and ∂_t^{-1} is the primitive defined by

$$\partial_t^{-1} \varphi(t, \mathbf{x}) = \int_0^t \varphi(s, \mathbf{x}) ds$$

Taking tangential traces on Σ , one then obtains systems of integral equations analogous to (D1)–(N4) for the unknown surface current and charge densities. Owing to special symmetries of the Maxwell equations, the set of four boundary integral operators V, K, K', W appearing in the boundary reduction of second-order problems is reduced to only two different boundary integral operators that we denote by V and K , defined by

$$\begin{aligned} V\varphi &= -\mathbf{n} \wedge S \left(\frac{1}{c} \partial_t \varphi \right) + \operatorname{curl}_\Gamma S(c \partial_t^{-1} \varphi) \\ K\varphi &= \frac{1}{2} (\gamma^+ + \gamma^-) \mathbf{n} \wedge \operatorname{curl} S(\varphi) \end{aligned}$$

In the definition of K , one takes the principal value that also corresponds to the mean value between the exterior trace γ^+ and the interior trace γ^- , analogous to the definition of the double-layer potential operator K in Section 2.2.1.

For the exterior initial value problem, the traces

$$\mathbf{v} = \mathbf{m} = \mathbf{n} \wedge \mathbf{E} \quad \text{and} \quad \varphi = \mu c \mathbf{j} = \sqrt{\frac{\mu}{\epsilon}} \mathbf{H} \wedge \mathbf{n}$$

then satisfy the two relations corresponding to the four integral equations (D1), (D2), (N1), (N2) of the *direct method*

$$\left(\frac{1}{2} - K \right) \mathbf{v} = -V\varphi \quad \text{and} \quad \left(\frac{1}{2} - K \right) \varphi = V\mathbf{v}$$

From a *single-layer representation*, that is, $[\mathbf{m}] = 0$ in the representation formula for the electric field, one obtains the time-dependent *electric field integral equation*, which can now be written as

$$V\varphi = \mathbf{g}$$

where \mathbf{g} is given by the tangential component of the incident field (see Chapter 26, this Volume).

2.3 Space-time variational formulations and Galerkin methods

We will not treat the analysis of second-kind boundary integral equations in detail here. Suffice it to say that the

key observation in the parabolic case is the fact that for smooth Γ , the operator norm in $L^p(\Sigma)$ of the weakly singular operator K tends to 0 as $T \rightarrow 0$. This implies that $(1/2) \pm K$ and $(1/2) \pm K'$ are isomorphisms in L^p (and also in C^∞), first for small T and then by iteration for all T . The operators K and K' being compact, one can use all the well-known numerical methods for classical Fredholm integral equations of the second kind, including Galerkin, collocation, Nyström methods (Pogorzelski, 1966; Kress, 1989), with the additional benefit that the integral equations are always uniquely solvable. If Γ has corners, these arguments break down, and quite different methods, including variational arguments, have to be used (Costabel, 1990; Dahlberg and Verchota, 1990; Brown, 1989; Brown and Shen, 1993; Adolfsson *et al.*, 1994).

2.3.1 Galerkin methods

For the first-kind integral equations, an analysis based on variational formulations is available. The corresponding numerical methods are space-time Galerkin methods. Their advantage is that they inherit directly the stability of the underlying variational method. In the elliptic case, this allows the well-known standard boundary element analysis of stability and errors, very similar to the standard finite element methods. In the parabolic case, the situation is still similar, but in the hyperbolic case, some price has to be paid for the application of 'elliptic' techniques. In particular, one has then to work with two different norms.

Let X be some Hilbert space and let a be a bilinear form on $X \times X$. If we assume that a is bounded on X :

$$\exists M : \forall u, v \in X : |a(u, v)| \leq M \|u\| \|v\|$$

but that a is elliptic only with respect to a smaller norm $\|\cdot\|_0$, associated with a space X_0 into which X is continuously embedded:

$$\exists \alpha > 0 : \forall u \in X : |a(u, u)| \geq \alpha \|u\|_0^2$$

then for the variational problem, find $u \in X$ such that

$$a(u, v) = (f, v) \quad \forall v \in X$$

and for its Galerkin approximation, find $u_N \in X_N$ such that

$$a(u_N, v_N) = (f, v_N) \quad \forall v_N \in X_N$$

there are stability and error estimates with a loss

$$\|u_N\|_0 \leq C \|u\|$$

$$\text{and} \quad \|u - u_N\|_0 \leq C \inf\{\|u - v_N\| \mid v_N \in X_N\}$$

The finite-dimensional space X_N for the Galerkin approximation of space-time integral equations is usually constructed as a tensor product of a standard boundary element space for the spatial discretization and of a space of one-dimensional finite element or spline functions on the interval $[0, T]$ for the time discretization. Basis functions are then of the form

$$\varphi_{ij}(t, x) = \chi_i(t)\psi_j(x) \quad (i = 1, \dots, I, \quad j = 1, \dots, J)$$

and the trial functions are of the form

$$u_N(t, x) = \sum_{i,j=1}^{I,J} \alpha_{ij} \varphi_{ij}(t, x)$$

The system of Galerkin equations for the unknown coefficients α_{ij} is

$$\sum_{i,j=1}^{I,J} \alpha(\varphi_{ij}, \varphi_{kl}) \alpha_{ij} = (f, \varphi_{kl}) \quad (k = 1, \dots, I, \quad l = 1, \dots, J)$$

In the following, we restrict the presentation to the single-layer potential operator V . We emphasize, however, that a completely analogous theory is available for the hypersingular operator W in all cases.

The variational methods for the first-kind integral operators are based on the first Green formula that gives, together with the jump relations, a formula valid again for all three types of equations: If φ and ψ are given on Γ or Σ , satisfy a finite number of conditions guaranteeing the convergence of the integrals on the right-hand side of the formula (2) below, and

$$u = S\varphi, \quad v = S\psi$$

then

$$\int_{\Gamma} \varphi \psi \, d\sigma = \int_{\mathbb{R}^n \setminus \Gamma} \{\nabla u \cdot \nabla v + u \Delta v\} \, dx \quad (2)$$

2.3.2 (E)

For the elliptic case, we obtain $(\langle \cdot, \cdot \rangle)_{\Gamma}$ denotes L^2 duality on Γ ;

$$(\varphi, V\psi)_{\Gamma} = \int_{\mathbb{R}^n \setminus \Gamma} (|\nabla u|^2 - \omega^2 |u|^2) \, dx$$

This gives the following theorem that serves as a model for the other two types. It holds not only for the simple case of the Laplacian but also, in particular its assertion

(ii), for more general second-order systems, including the Lamé system of linear elasticity (Costabel, 1988).

Theorem 1. Let Γ be a bounded Lipschitz surface, open or closed. $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$ denote the usual Sobolev spaces, and $\tilde{H}^{-1/2}(\Gamma)$ for an open surface is the dual of $H^{1/2}(\Gamma)$. Then

(i) For $\omega = 0$, $n \geq 3$: $V: \tilde{H}^{-1/2}(\Gamma) \rightarrow H^{1/2}(\Gamma)$ is an isomorphism, and there is an $\alpha > 0$ such that

$$(\varphi, V\varphi)_{\Gamma} \geq \alpha \|\varphi\|_{\tilde{H}^{-1/2}(\Gamma)}^2$$

(ii) For any ω and n , there is an $\alpha > 0$ and a compact quadratic form k on $\tilde{H}^{-1/2}(\Gamma)$ such that

$$\operatorname{Re}(\varphi, V\varphi)_{\Gamma} \geq \alpha \|\varphi\|_{\tilde{H}^{-1/2}(\Gamma)}^2 - k(\varphi)$$

(iii) If ω is not an interior or exterior eigenfrequency, then V is an isomorphism, and every Galerkin method in $\tilde{H}^{-1/2}(\Gamma)$ for the equation $V\psi = g$ is stable and convergent.

2.3.3 (P)

For the parabolic case of the heat equation, integration over t in the Green formula (2) gives

$$\begin{aligned} (\varphi, V\psi)_{\Sigma} &= \int_0^T \int_{\mathbb{R}^n \setminus \Gamma} \{|\nabla_x u(t, x)|^2 + \partial_t u \bar{u}\} \, dx \, dt \\ &= \iint_{\Sigma} |\nabla_x u(t, x)|^2 \, dx \, dt + \frac{1}{2} \int_{\mathbb{R}^n} |u(T, x)|^2 \, dx \end{aligned}$$

From this, the positivity of the quadratic form associated with the operator V is evident. What is less evident is the nature of the energy norm for V , however. It turns out (Arnold and Noon, 1989; Costabel, 1990) that one has to consider anisotropic Sobolev spaces of the following form

$$\tilde{H}_0^{\alpha, \beta}(\Sigma) = L^2(0, T; \tilde{H}^{\alpha}(\Gamma)) \cap H_0^{\beta}(0, T; L^2(\Gamma))$$

The index 0 indicates that zero initial conditions at $t = 0$ are incorporated. The optional \sim means zero boundary values on the boundary of the (open) manifold Γ . One has the following theorem, which is actually simpler than its elliptic counterpart because the operators are always invertible due to their Volterra nature.

Theorem 2. Let Γ be a bounded Lipschitz surface, open or closed, $n \geq 2$.

(i) $V: \tilde{H}_0^{(-1/2), (-1/4)}(\Sigma) \rightarrow H_0^{(1/2), (1/4)}(\Sigma)$ is an isomorphism, and there is an $\alpha > 0$ such that

$$(\varphi, V\varphi)_{\Sigma} \geq \alpha \|\varphi\|_{\tilde{H}_0^{(-1/2), (-1/4)}(\Sigma)}^2$$

(ii) Every Galerkin method in $\tilde{H}_0^{(-1/2), (-1/4)}(\Sigma)$ for the equation $V\psi = g$ converges. The Galerkin matrices have positive-definite symmetric part. Typical error estimates are of the form

$$\begin{aligned} \|\varphi - \varphi_{h,k}\|_{(-1/2), (-1/4)} \\ \leq C(h^{r+(1/2)} + k^{r/2+(1/4)}) \|\varphi\|_{r, r/2} \end{aligned}$$

if $\varphi_{h,k}$ is the Galerkin solution in a tensor product space of splines of mesh-size k in time and finite elements of mesh-size h in space.

2.3.4 (H)

For the wave equation, choosing $\varphi = \bar{\psi}$ in the Green formula (2) does not give a positive-definite expression. Instead, one can choose $\varphi = \partial_t \bar{\psi}$. This corresponds to the usual procedure for getting energy estimates in the weak formulation of the wave equation itself where one uses $\partial_t u$ as a test function, and it gives

$$\begin{aligned} (\partial_t \varphi, V\psi)_{\Sigma} &= \int_0^T \int_{\mathbb{R}^n \setminus \Gamma} \{\partial_t \nabla_x \bar{u} \cdot \nabla_x u + \partial_t \bar{u} \partial_t^2 u\} \, dx \, dt \\ &= \frac{1}{2} \int_{\mathbb{R}^n \setminus \Gamma} (|\nabla_x u(T, x)|^2 + |\partial_t u(T, x)|^2) \, dx \end{aligned}$$

Once again, as in the elliptic case, this shows the close relation of the operator V with the total energy of the system. In order to obtain a norm $(H^1(Q))$ on the right-hand side, one can integrate a second time over t . But in any case, here the bilinear form $(\partial_t \varphi, V\psi)_{\Sigma}$ will not be bounded in the same norm where its real part is positive. So there will be a loss of regularity, and any error estimate has to use two different norms. No 'natural' energy space for the operator V presents itself.

2.4 Fourier-Laplace analysis and Galerkin methods

A closer view of what is going on can be obtained using space-time Fourier transformation. For this, one has to assume that Γ is flat, that is, a subset of \mathbb{R}^{n-1} . Then all the operators are convolutions and as such are represented by multiplication operators in Fourier space. If Γ is not flat but smooth, then the results for the flat case describe the principal part of the operators. To construct a complete analysis, one has to consider lower order terms coming from coordinate transformations and localizations. Whereas this is a well-known technique in the elliptic and parabolic cases, namely, part of the calculus of pseudodifferential

operators, it has so far prevented the construction of a completely satisfactory theory for the hyperbolic case.

We denote the dual variables to (t, x) by (ω, ξ) , and x' and ξ' are the variables related to $\Gamma \subset \mathbb{R}^{n-1}$. It is then easily seen that the form of the single-layer potential is

$$\widehat{V\psi}(\omega, \xi') = \frac{1}{2}(|\xi'|^2 - \omega^2)^{-1/2} \hat{\psi}(\omega, \xi') \quad (\mathcal{E})$$

$$\widehat{V\psi}(\omega, \xi') = \frac{1}{2}(|\xi'|^2 - i\omega)^{-1/2} \hat{\psi}(\omega, \xi') \quad (\mathcal{P})$$

$$\widehat{V\psi}(\omega, \xi') = \frac{1}{2}(|\xi'|^2 - \omega^2)^{-1/2} \hat{\psi}(\omega, \xi') \quad (\mathcal{H})$$

Note that (\mathcal{E}) and (\mathcal{H}) differ only in the role of ω ; for (\mathcal{E}) , it is a fixed parameter, for (\mathcal{H}) , it is one of the variables, and this is crucial in the application of Parseval's formula for $(\varphi, V\varphi)$.

2.4.1 (E)

For the elliptic case, the preceding formula implies Theorem 1: If $\omega = 0$, then the function $(1/2)|\xi'|^{-1}$ is positive and for large $|\xi'|$ equivalent to $(1 + |\xi'|^2)^{-1/2}$, the Fourier weight defining the Sobolev space $H^{-1/2}(\Gamma)$. If $\omega \neq 0$, then the principal part (as $|\xi'| \rightarrow \infty$) is still $(1/2)|\xi'|^{-1}$, so only a compact perturbation is added. There is an additional observation by Ha-Duong (1990): If ω is real, then $(1/2)(|\xi'|^2 - \omega^2)^{-1/2}$ is either positive or imaginary, so its real part is positive except on the bounded set $|\xi'| \leq |\omega|$. This implies

Proposition 1. Let $\omega^2 > 0$, Γ flat, supp φ compact. Then there is an $\alpha(\omega) > 0$ such that

$$\operatorname{Re}(\varphi, V\varphi)_{\Gamma} \geq \alpha(\omega) \|\varphi\|_{\tilde{H}^{-1/2}}^2$$

The work of transforming this estimate into error estimates for the BEM in the hyperbolic case is still incomplete. See Ha-Duong (2003) for a review of the state of the art on this question.

2.4.2 (P)

For the parabolic case, the symbol of the single-layer potential,

$$\sigma_V(\omega, \xi') = \frac{1}{2}(|\xi'|^2 - i\omega)^{-1/2}$$

again has a positive real part. In addition, it is sectorial:

$$|\arg \sigma_V(\omega, \xi')| \leq \frac{\pi}{4}$$

This has the consequence that its real part and absolute value are equivalent (an 'elliptic' situation):

$$C_1 ||\xi'|^2 - i\omega|^{-1/2} \leq \operatorname{Re} \sigma_V(\omega, \xi') \leq C_2 ||\xi'|^2 - i\omega|^{-1/2}$$

In addition, for large $|\xi|^2 + |\omega|$, this is equivalent to $((1 + |\xi|^2) + |\omega|)^{-1/2}$, the Fourier weight defining the space $H^{(-1/2), (-1/4)}(\Sigma)$. This explains Theorem 2. It also clearly shows the difference between the single-layer heat potential operator on the boundary and the heat operator $\partial_t - \Delta$ itself. The symbol of the latter is $|\xi|^2 - i\omega$, and the real part $|\xi|^2$ and the absolute value $(|\xi|^2 + |\omega|)^{1/2}$ of this function are not equivalent uniformly in ξ and ω .

2.4.3 (\mathcal{H})

In the hyperbolic case, the symbol σ_V does not have a positive real part. Instead, one has to multiply it by $i\bar{\omega}$ and to use a complex frequency $\omega = \omega_R + i\omega_I$ with $\omega_I > 0$ fixed. Then, one gets

$$\operatorname{Re}(i\bar{\omega}(|\xi|^2 - \omega^2)^{1/2}) \geq \frac{\omega_I}{2}(|\xi|^2 + |\omega|)^{1/2}$$

and similar estimates given first by Bamberger and Ha Duong (1986). Note that with respect to $|\omega|$, one is losing an order of growth here. For fixed ω_I , the left-hand side is bounded by $|\omega|^2$, whereas the right-hand side is $\mathcal{O}(|\omega|)$. One introduces another class of anisotropic Sobolev spaces of the form

$$H^{s,r}(\mathbb{R} \times \Gamma) = \{u \mid u, \partial_t^r u \in H^s(\mathbb{R} \times \Gamma)\}$$

with the norm

$$\|u\|_{s,r,\omega_I} = \int_{\operatorname{Im} \omega = \omega_I} \int_{\mathbb{R}^{n-1}} |\omega|^{2r} (|\xi|^2 + |\omega|^2)^s |\hat{u}(\omega, \xi')|^2 d\xi' d\omega$$

We give one example of a theorem obtained in this way.

Theorem 3. Let Γ be bounded and smooth, $r, s \in \mathbb{R}$. Then

- (i) $V: \tilde{H}_0^{s,r+1}(\Sigma) \rightarrow H_0^{s+1,r}(\Sigma)$ and $V^{-1}: H^{s+1,r+1}(\Sigma) \rightarrow \tilde{H}_0^{s,r}(\Sigma)$ are continuous.
- (ii) Let $\omega_I > 0$ and the bilinear form $a(\varphi, \psi)$ be defined by

$$a(\varphi, \psi) = \int_0^\infty e^{-2\omega_I t} \int_\Gamma (V\varphi)(t, x) \overline{\partial_t \psi}(t, x) d\alpha(x) dt$$

Then there is an $\alpha > 0$ such that

$$\operatorname{Re} a(\varphi, \varphi) \geq \alpha \|\varphi\|_{(-1/2), 0, \omega_I}^2$$

- (iii) The Galerkin matrices for the scheme: Find $\varphi_N \in X_N$ such that

$$a(\varphi_N, \psi) = (g, \partial_t \psi)_\Sigma \quad \forall \psi \in X_N$$

have positive-definite hermitian part, and there is an error estimate

$$\|\varphi - \varphi_N\|_{(-1/2), 0, \omega_I} \leq C \omega_I^{-1/2} \inf_{\psi \in X_N} \|\varphi - \psi\|_{(-1/2), 1, \omega_I}$$

Thus, one has unconditional stability and convergence for $\omega_I > 0$. In practical computations, one will use the bilinear form $a(\varphi, \psi)$ for $\omega_I = 0$, where the error estimate is no longer valid. Instabilities have been observed that are, however, probably unrelated to the omission of the exponential factor. They are also not caused by a too large CFL number (ratio between time step and spatial mesh width). In fact, too small and too large time steps have both been reported to lead to instabilities.

Corresponding results for elastodynamics and for electro-dynamics can be found in the literature (besides the above-mentioned works, see the references given in Chudinovich, 2001 and in Bachelot *et al.*, 2001).

2.5 Collocation methods

In order to avoid the high-dimensional integrations necessary for the computation of the matrix elements in a Galerkin method such as the ones described in Theorems 2 and 3, one often uses collocation methods. Just like in the elliptic case, even for the classical first-kind integral operators for the Laplace operator, the mathematical analysis lags seriously behind the practical experiences.

In more than two dimensions, only for very special geometries that are amenable to Fourier analysis, stability of collocation schemes can be shown. For time-dependent integral equations, even two-space dimensions create problems that only recently have been overcome, and this is only for special geometries, mainly flat boundaries or toroidal boundaries.

Collocation schemes for the single-layer potential integral equation (D3) are easy to formulate. One usually takes basis functions of tensor product form, that is,

$$\varphi_{ij}(t, x) = \chi_i(t) \psi_j(x)$$

where χ_i ($i = 1, \dots, M$) is a basis of a space of finite elements (splines) of degree d_t on the interval $[0, T]$, and ψ_j ($j = 1, \dots, N$) is a basis of a space of finite elements of degree d_x on the boundary Γ . Then, the trial functions are of the form

$$u_{kh}(t, x) = \sum_{i,j=1}^{M,N} \alpha_{ij} \varphi_{ij}(t, x)$$

Here, the indices kh indicate the time step $k \sim T/M$ and the mesh width h of the discretization of the boundary Γ .

The linear system for the unknown coefficients is obtained from the equations

$$Vu_{kh}(t_i, x_j) = g(t_i, x_j)$$

where $t_i \in [0, T]$ ($i = 1, \dots, M$) are the time collocation points and $x_j \in \Gamma$ ($j = 1, \dots, N$) are the space collocation points. The collocation points are usually chosen in a 'natural' way, meaning midpoints for even degree splines in time, nodes for odd degree splines in time, barycenters for piecewise constants $d_x = 0$, nodes of the finite element mesh on Γ for $d_x = 1$, and, more generally, nodes of suitable quadrature rules for other values of d_x .

2.5.1 (\mathcal{P})

For the heat equation in a smooth domain in two-space dimensions, it was shown in Costabel and Saramen (2000, 2003) that for $d_x = 0, 1$, one gets convergence in anisotropic Sobolev spaces of the 'parabolic' class defined in subsection 2.3.3. There is a condition for optimality of the convergence that corresponds to a kind of anisotropic quasiniformity:

$$k \sim h^2$$

2.5.2 (\mathcal{H})

For the retarded potential integral equation, that is, the equation of the single-layer potential for the wave equation in three-space dimensions, Davies and Duncan (2003) prove rather complete stability and convergence results for the case of a flat boundary.

3 LAPLACE TRANSFORM METHODS

To pass from the time domain to the frequency domain, we define the (Fourier-) Laplace transform by

$$\hat{u}(\omega) = \mathcal{L}u(\omega) = \int_0^\infty e^{-i\omega t} u(t) dt \quad (3)$$

If u is integrable with a polynomial weight or, more generally, a tempered distribution, and if, as we assume here throughout, $u(t) = 0$ for $t < 0$, then \hat{u} is holomorphic in the upper half plane $\{\omega = \omega_R + i\omega_I \mid \omega_R \in \mathbb{R}, \omega_I > 0\}$. The inversion formula is

$$u(t) = \mathcal{L}^{-1}u(t) = \frac{1}{2\pi} \int_{-\infty-i\omega_I}^{\infty-i\omega_I} e^{-i\omega t} \hat{u}(\omega) d\omega \quad (4)$$

Frequently, it is customary to define the Laplace integral by

$$\int_0^\infty e^{-st} u(t) dt$$

which is the same as (3) when s and ω are related by $s = -i\omega$. The upper half plane $\omega_I > 0$ coincides with the right half plane $\operatorname{Re} s > 0$.

The function $t \mapsto u(t)$ can take values in some Banach space (Arendt *et al.*, 2001), for example, in a space of functions depending on x , in which case we write

$$\hat{u}(\omega, x) = \mathcal{L}u(\omega, x) = \int_0^\infty e^{-i\omega t} u(t, x) dt$$

By Laplace transformation, both the parabolic and the hyperbolic initial-boundary value problems are transformed into elliptic boundary value problems with an eigenvalue parameter λ depending on the frequency ω . Thus, both the heat equation $(\partial_t - \Delta)u = 0$ and the wave equation $(c^{-2}\partial_t^2 - \Delta)u = 0$ are transformed into the Helmholtz equation $(\Delta - \lambda)\hat{u}(\omega, x) = 0$, where

$$\lambda(\omega) = -i\omega \text{ for the heat equation}$$

$$\text{and } \lambda(\omega) = -\frac{\omega^2}{c^2} \text{ for the wave equation}$$

The idea of the Laplace transform boundary integral equation method is to solve these elliptic boundary value problems for a finite number of frequencies with a standard boundary element method and then to insert the results into a numerical approximation of the Laplace inversion integral (4).

There exist various algorithms for numerical inverse Laplace transforms (see e.g. Davies and Martin, 1979 or Abate and Whitt, 1995). One will, in general, first replace the line of integration $\{\operatorname{Im} \omega = \omega_I\}$ by a suitable equivalent contour \mathcal{C} and then choose some quadrature rule approximation of the integral. The end result will be of the form

$$u(t) = \frac{1}{2\pi} \int_{\mathcal{C}} e^{-i\omega t} \hat{u}(\omega) d\omega \sim \sum_{\ell=1}^L w_\ell e^{-i\omega_\ell t} \hat{u}(\omega_\ell) \quad (5)$$

with quadrature weights w_ℓ and a finite number of frequencies ω_ℓ .

One obvious candidate for such a quadrature formula is the trapezoidal rule on a large interval $[-R, R]$, where the line $\{\operatorname{Im} \omega = \omega_I\}$ is replaced by $[-R, R] + i\omega_I$. This can then be evaluated by fast Fourier transform, which is clear when we write the Laplace inversion integral as inverse

Fourier transform over the real line:

$$u(t) = \mathcal{L}^{-1}\hat{u}(t) = e^{i\omega t} \mathcal{F}_{\omega \mapsto i\omega}^{-1}[\hat{u}(\omega, +i\omega_t)]$$

Let us describe the resulting procedure in more detail for the formulation with a single-layer potential representation for the initial-Dirichlet problem, keeping in mind that any other type of boundary integral equation constructed in Section 2.2.2 would do as well and lead to a similar formalism. By Laplace transform, we get the boundary value problem

$$\begin{aligned} (\Delta - \lambda(\omega))\hat{u}(\omega, x) &= 0 \quad \text{in } \Omega \\ \hat{u}(\omega, x) &= \hat{g}(\omega, x) \quad \text{on } \Gamma \end{aligned}$$

where the right-hand side is the Laplace transform of the given boundary data g . For the unknown density ψ , we get the first-kind integral equation on Γ

$$V_{\lambda(\omega)}\hat{\psi}(\omega) = \hat{g}(\omega) \quad (6)$$

where $V_{\lambda(\omega)}$ is the weakly singular integral operator generated by convolution with the kernel (in three dimensions)

$$G_{\lambda(\omega)}(x) = \frac{e^{\sqrt{\lambda(\omega)}|x|}}{4\pi|x|}$$

Now, let $V_{\lambda(\omega),A}$ be some finite-dimensional boundary element approximation of $V_{\lambda(\omega)}$, so that

$$\hat{\psi}_A(\omega) = V_{\lambda(\omega),A}^{-1}\hat{g}(\omega)$$

is the corresponding approximate solution of equation (6). Inserting this into the numerical inversion formula (5) finally gives the following expression for the approximation of the unknown density $\psi(t, x)$ via the Laplace transform boundary element method

$$\psi_A(t, x) = \sum_{l=1}^L w_l e^{-i\omega_l t} \left(V_{\lambda(\omega_l),A}^{-1} \hat{g}(\omega_l) \right) (x) \quad (7)$$

Note that on this level of abstraction, formula (7) looks the same for the parabolic case of the heat equation, the hyperbolic case of the wave equation, or even the dissipative wave equation. The only difference is the function $\lambda(\omega)$, which then determines, depending on the contour C and its discretization ω_l , for which complex frequencies $\sqrt{(-\lambda(\omega_l))}$ the single-layer potential operator has to be numerically inverted.

For practical computations, this difference can be essential. In a precise quadrature rule in (5), which is needed for high resolution in time, there will be some ω_l with

large absolute values. In the hyperbolic case (but not in the parabolic case!), this means large negative real parts for $\lambda(\omega_l)$, hence highly oscillating kernels, and some machinery for high-frequency boundary element methods has to be put in place (see e.g. Bruno, 2003).

Applications of the Laplace transform boundary integral equation methods in elastodynamics have a long history (Cruse and Rizzo, 1968; Cruse, 1968). For generalizations such as viscoelasticity, poroelasticity, or piezoelectricity, these methods are more practical than the space-time boundary integral equation methods because space-time fundamental solutions are not explicitly known or very complicated (Gaul and Schanz, 1999; Schanz, 1999; Schanz, 2001a; Wang *et al.*, 2003). Recently, Laplace domain methods related to the operational quadrature method (see subsection 4.4) have been used successfully in practice (Schanz and Antes, 1997a, b; Schanz, 2001b; Telles and Vera-Tudela, 2003).

A final remark on the Laplace transform boundary element method: Instead of, as described in this section, performing first the Laplace transform and then the reduction to the boundary, one can also first construct the space-time boundary integral equations as described in the previous section and then apply the Laplace transform. It is easy to see that the resulting frequency-domain boundary integral equations are exactly the same in both procedures.

4 TIME-STEPPING METHODS

In the previous sections, the boundary reduction step was performed before any discretization had taken place. In particular, the description of the transient behavior of the solution by a finite number of degrees of freedom was introduced via a Galerkin or collocation method for the space-time integral equation or via numerical Laplace inversion, only after the construction of the boundary integral equation.

It is possible to invert the order of these steps by first applying a time-discretization scheme to the original initial-boundary value problem and then using a boundary integral equation method on the resulting problem that is discrete in time and continuous in space. One advantage of this idea is similar to the motivation of the Laplace transform method: The parabolic and hyperbolic problems are reduced to elliptic problems for which boundary element techniques are well known. Another attraction is the idea that once a procedure for one time step is constructed, one can march arbitrarily far in time by simply repeating this same procedure.

In this section, we will, for simplicity, only treat the parabolic case of the initial-Dirichlet problem for the heat

equation. Quite analogous procedures are also possible for the hyperbolic case, and, in particular, the operational quadrature method has been analyzed for both the parabolic and the hyperbolic situation (see Lubich, 1994). In the literature on applied boundary element methods, one can find many successful applications of similar time-stepping schemes to parabolic and hyperbolic problems of heat transfer, fluid dynamics, elastodynamics, and various generalizations (Nardini and Brebbia, 1983; Partridge *et al.*, 1992; Gaul *et al.*, 2003).

4.1 Time discretization

We consider the initial-boundary value problem

$$\begin{aligned} (\partial_t - \Delta)u(t, x) &= 0 \quad \text{in } Q \\ u &= g \quad \text{on } \Sigma \\ u(t, x) &= 0 \quad \text{for } t \leq 0 \end{aligned} \quad (8)$$

as an ordinary differential equation in time with operator coefficients. Consequently, we can employ any kind of one-step or multistep method known from the numerical analysis of ordinary differential equations. Only implicit schemes are of interest here, for two reasons: The first reason is the stability of the resulting scheme, and, secondly, explicit schemes would not really require a boundary integral equation method.

The solution $u(t, x)$ for $0 \leq t \leq T$ is approximated by a sequence $u^n(x)$, $n = 0, \dots, N$, where

$$\begin{aligned} u^n &\text{ is understood as an approximation of } u(t_n, \cdot), \\ t_n &= nk = \frac{nT}{N} \end{aligned}$$

The simplest discretization of the derivative ∂_t with time step k is the backward difference that gives the backward Euler scheme for (8)

$$\begin{aligned} \frac{u^n - u^{n-1}}{k} - \Delta u^n &= 0 \quad \text{in } \Omega \quad (n = 1, \dots, N) \\ u^n(x) &= g^n(x) = g(t_n, x) \quad \text{on } \Gamma \quad (n = 1, \dots, N) \\ u^0 &= 0 \quad \text{for } t \leq 0 \end{aligned} \quad (9)$$

The actual elliptic boundary value problem that one has to solve at each time step, $n = 1, \dots, N$ is therefore

$$u^n - k \Delta u^n = u^{n-1} \quad \text{in } \Omega; \quad u^n = g^n \quad \text{on } \Gamma \quad (10)$$

Higher-order approximation in time can be achieved by multistep methods of the form

$$\sum_{j=0}^r \alpha_j u^{n-j} - k \sum_{j=0}^r \beta_j \Delta u^{n-j} = 0 \quad \text{in } \Omega; \quad u^n = g^n \quad \text{on } \Gamma \quad (11)$$

The coefficients α_j and β_j define the characteristic function of the multistep scheme

$$\delta(\zeta) = \frac{\sum_{j=0}^r \alpha_j \zeta^j}{\sum_{j=0}^r \beta_j \zeta^j}$$

Consistency of the scheme (11) is characterized by $\delta(1) = 0$, $\delta'(1) = -1$, and the scheme is accurate of order p if $\delta(e^{-\zeta})/\zeta = 1 + \mathcal{O}(\zeta^p)$ as $\zeta \rightarrow 0$. One can assume that $\alpha_0 \beta_0 > 0$.

4.2 One step at a time

The problem to solve for one time step in both (10) and (11) is of the form

$$\eta^2 u - \Delta u = f \quad \text{in } \Omega; \quad u = g \quad \text{on } \Gamma \quad (12)$$

Here, $\eta^2 = 1/k$ for (10) and $\eta^2 = \alpha_0/(k\beta_0)$ for (11). The right-hand side f is computed from the solution of the previous time step(s), and it has no reason to vanish except possibly for the very first time step. For the integral equation method, we therefore have to apply a representation formula that takes into account this inhomogeneous differential equation.

Let $u_p = Pf$ be a particular solution of the equation $\eta^2 u_p - \Delta u_p = f$ in Ω . Then, $u_0 = u - u_p$ satisfies the homogeneous equation and can therefore be computed by a standard boundary integral equation method, for example, by one of the methods from Section 2.2. For an exterior domain, we thus have the representation formula in Ω

$$\begin{aligned} u_0(x) &= \int_{\Gamma} \{\partial_{n(y)} G(x-y) u_0(y) - G(x-y) \partial_n u_0(y)\} d\sigma(y) \\ &= \mathcal{D}(\gamma_0 u_0)(x) - \mathcal{S}(\gamma_1 u_0)(x) \end{aligned}$$

Here, G is the fundamental solution of the Helmholtz equation given in the three-dimensional case by

$$G(x) = \frac{e^{-\eta|x|}}{4\pi|x|}$$

Using our abbreviations for the single- and double-layer potentials and

$$\gamma_0 u = u|_{\Gamma}, \quad \gamma_1 u = \partial_n u|_{\Gamma}$$

we have the representation for u

$$u = \mathcal{D}(\gamma_0 u) - \mathcal{S}(\gamma_1 u) + Pf - \mathcal{D}(\gamma_0 Pf) + \mathcal{S}(\gamma_1 Pf) \quad (13)$$

For the unknown $\varphi = \gamma_1 u$ in the direct method for the Dirichlet problem or for the unknown ψ in a single-layer potential representation

$$u = \mathcal{S}\psi + Pf \quad (14)$$

or the unknown w in a double-layer representation

$$u = \mathcal{D}w + Pf \quad (15)$$

this leads to the choice of integral equations

$$V\psi = (-\frac{1}{2} + K)g + (\frac{1}{2} - K)\gamma_0 Pf + V\gamma_1 Pf \quad (D1)$$

$$(\frac{1}{2} + K')\varphi = -Wg + W\gamma_0 Pf + (-\frac{1}{2} + K')\gamma_1 Pf \quad (D2)$$

$$V\psi = g - \gamma_0 Pf \quad (D3)$$

$$(\frac{1}{2} + K)w = g - \gamma_0 Pf \quad (D4)$$

These are standard boundary integral equations that can be discretized and numerically solved in many ways. The peculiarity is the appearance of the particular solution Pf in the representation formula (13) and in the integral equations (D1)–(D4).

There are various possibilities for the construction of (an approximation of) Pf . Let us mention some of them that are being used in the boundary element literature and practice.

4.2.1 Newton potential

In the standard representation formula for the inhomogeneous Helmholtz equation derived from Green's formula, Pf appears in the form

$$Pf(x) = \int_{\Omega} G(x-y) f(y) dy$$

This representation has the advantage that the last two terms in the representation formula (13) cancel, and therefore also the integral equations (D1) and (D2) simplify in that the integral operators acting on the traces of Pf are absent.

For computing the Newton potential, the domain Ω has to be discretized, thus neutralizing one of the advantages of the boundary element method, namely the reduction of the dimension. Note, however, that this domain discretization is done only for purposes of numerical integration. No finite element grid has to be constructed. It is also to be noted that the domain discretization only enters into the computation of the right-hand side; the size of the linear system to be solved is not affected.

4.2.2 Fourier series

Another method to get an approximate particular solution Pf is to embed the domain Ω into a rectangular domain $\tilde{\Omega}$, then approximate an extension of f to $\tilde{\Omega}$ by trigonometric polynomials using fast Fourier transform, solve the Helmholtz equation in Fourier space, and go back by FFT again. Other fast Helmholtz solvers that exist for simple domains can be used in the same way.

4.2.3 Radial basis functions

In the previous subsections, the right-hand side f was approximated by a linear combination of special functions for which particular solutions of the Helmholtz equation are known: the Dirac distribution for the Newton potential method, and exponential functions for the FFT method. The particular solution Pf is then given by the corresponding linear combination of the individual particular solutions. Other special functions that can serve in the same way are radial basis functions, in the simplest case functions of the form $|x - x_j|$, where the x_j belong to some discretization of Ω by an unstructured grid. One advantage of the radial basis function technique is that there exist many practical and theoretical results about interpolation by such functions (Powell, 1992; Faul and Powell, 1999).

4.2.4 Higher fundamental solutions

In the first time step, the solution $u = u^1$ is given, after solving the appropriate boundary integral equations, by the representation formula (13) with $f = 0$, that is, by a combination of single- and double-layer potentials. This u^1 is then used as right-hand side f in the next time step. A particular solution Pf can then be found, without any domain integral, by replacing the fundamental solution G of $(\eta^2 - \Delta)$ in the representation formula by a fundamental solution $G^{(1)}$ of $(\eta^2 - \Delta)^2$ satisfying

$$(\eta^2 - \Delta)G^{(1)}(x) = G(x)$$

Thus, if

$$f(x) = \int_{\Gamma} \{\partial_n G(x-y)w(y) - G(x-y)\varphi\} d\sigma(y)$$

then a particular solution Pf is given by

$$Pf(x) = \int_{\Gamma} \{\partial_n G^{(1)}(x-y)w(y) - G^{(1)}(x-y)\varphi\} d\sigma(y)$$

In the next time step, the right-hand side is then constructed from single- and double-layer potentials plus this Pf . Repeating the argument, one obtains a particular solution by using a fundamental solution $G^{(n)}$ of $(\eta^2 - \Delta)^n$. In the n th time step, one then needs to use higher-order fundamental solutions $G^{(j)}$, $(j < n)$, which satisfy the recurrence relations

$$(\eta^2 - \Delta)G^{(j+1)}(x) = G^{(j)}(x)$$

Such functions $G^{(j)}$ can be given explicitly in terms of Bessel functions. In this way, the whole time-marching scheme can be performed purely on the boundary, without using domain integrals or any other algorithm requiring discretization of the domain Ω . Two other points of view that can lead, eventually, to an entirely equivalent algorithm for the time-discretized problem, are described in the following sections.

4.3 All time steps at once

Just as in the construction of the space-time integral equations the heat equation or wave equation was not considered as an evolution equation, that is, as an ordinary differential equation with operator coefficients, but as a translation invariant operator on \mathbb{R}^{1+n} whose fundamental solution was used for integral representations, one can consider the time-discretized problem as a translation invariant problem on $\mathbb{Z} \times \mathbb{R}^n$ and construct a space-time fundamental solution for this semidiscretized problem. The role of the time derivative is then played by its one-step or multistep discretization, as in (10) or (11), and the role of the inverse of the time derivative and of other finite time convolutions appearing in the space-time integral operators is played by finite discrete convolutions.

In simple cases, such discrete convolution operators can be inverted explicitly. For a two-part recurrence relation, such as the backward Euler method (9), the convolution operator can be represented by a triangular Toeplitz matrix with just one lower side diagonal. Let u denote the vector u^1, \dots, u^N and define g correspondingly. Then the backward Euler scheme (10) can be written as a system

$$Au = 0 \quad \text{in } \Omega; \quad u = g \quad \text{on } \Gamma \quad (16)$$

Here, A is an elliptic system of second order, given by the matrix elements

$$a_{i,j} = 1 - k\Delta; \quad a_{i,j-1} = -1; \quad \text{all other } a_{i,j} = 0$$

Once a fundamental solution of this system is found, the system of equations (16) can be solved numerically by standard elliptic boundary element methods. Because of the simple form of A , such a fundamental solution can be written using the higher fundamental solutions $G^{(j)}$ of the Helmholtz equation defined in Section 4.2.4. It is a lower triangular Toeplitz matrix \mathcal{G} with entries $(G_{i,j})$, where

$$\begin{aligned} G_{i,j}(x) &= G(x) \\ G_{i,j}(x) &= G^{(i-j)}(x) \quad \text{for } j < i \\ G_{i,j}(x) &= 0 \quad \text{for } j < i \end{aligned}$$

All boundary integral operators constructed from this fundamental solution will have the same lower triangular Toeplitz (finite convolution) structure, and their solutions can be found by inverting the single operator that generates the diagonal and by subsequent back substitution.

For a detailed description of the approximation of the two-dimensional initial-Dirichlet problem for the heat equation using such a method, including formulas for the kernels in \mathcal{G} and a complete error analysis of the resulting second-kind integral equation as well as numerical results, see Chapko and Kress (1997).

4.4 The operational quadrature method

In the previous section, the simple structure of the backward Euler scheme was essential. The resulting numerical approximation is of only first order in time. If one wants to use schemes that are of higher order in time, one can employ multistep methods, as described above. The resulting schemes still have the lower triangular Toeplitz structure of finite discrete convolutions in time. From the algebraic structure of these convolutions, it is clear that fundamental solutions, the resulting boundary integral operators, and their solution operators all have this finite convolution structure.

Explicit constructions of kernels, however, will not be possible, in general. Just as for the original continuous-time problem the appropriate functional transform – the Laplace transform – allowed the reduction of the parabolic to elliptic problems, here, for the discrete-time problem, one can use the appropriate functional transform – namely the z -transform. In order to conserve the approximation order of the multistep method, one has to use a certain

translation between continuous convolutions and discrete convolutions, or, equivalently, between Laplace transforms and z -transforms.

For the generators of the convolution algebras, namely, the derivative ∂_t in the continuous case and its time step k discretization ∂_t^k , this translation is given by the definition (11) of the multistep method, characterized by the rational function $\delta(z)$. For the whole convolution algebras, this translation leads to the discretization method described by Lubich's *operational quadrature* method, see Lubich and Schneider (1992) and Lubich (1994). The general translation rule is the following (we use our notation for the (Fourier-)Laplace transform introduced above, not Lubich's notation):

Denote a finite convolution operator with operator-valued coefficients by

$$\widehat{K}(i\partial_t)u(t) = \mathcal{L}_{\omega \mapsto t}^{-1}[\widehat{K}(\omega)\widehat{u}(\omega)]$$

If $\widehat{K}(\omega)$ decays sufficiently rapidly in the upper half plane, this operator is given by an integrable kernel K whose Laplace transform is $\widehat{K}(\omega)$:

$$\widehat{K}(i\partial_t)u(t) = \int_0^t K(s)u(t-s)ds$$

The corresponding discrete convolution operator is given by

$$(\widehat{K}(i\partial_t^k)u)_n = \sum_{j=0}^n K_j u_{n-j}$$

where the coefficients K_j are defined by their z -transform

$$\sum_{j=0}^{\infty} K_j z^j = \widehat{K}\left(i\frac{\delta(z)}{k}\right)$$

Here, k is the time step and $\delta(z)$ is the characteristic function of the multistep method. The inverse of the z -transform is given by the Cauchy integral over some circle $|z| = \rho$

$$K_j = \frac{1}{2\pi i} \int_{|z|=\rho} \widehat{K}\left(i\frac{\delta(z)}{k}\right) z^{-j-1} dz$$

It is not hard to see that this translation rule reduces, for the case of the derivative $\partial_t = \widehat{K}(i\partial_t)$ with $\widehat{K}(\omega) = -i\omega$, to the convolution defined by the characteristic function $\delta(z)$:

$$\partial_t^k u_n = \sum_{j=0}^n \delta_j u_{n-j} \quad \text{with} \quad \delta(z) = \sum_{j=0}^{\infty} \delta_j z^j$$

In addition, this translation rule is an algebra homomorphism, that is, it respects compositions of (operator-valued) convolution operators. This is easy to see because

$$\begin{aligned} \widehat{K}_1(i\partial_t)\widehat{K}_2(i\partial_t) &= (\widehat{K}_1\widehat{K}_2)(i\partial_t) \\ \text{and also} \quad \widehat{K}_1(i\partial_t^k)\widehat{K}_2(i\partial_t^k) &= (\widehat{K}_1\widehat{K}_2)(i\partial_t^k) \end{aligned}$$

By the relation $z = e^{i\omega k}$, one can see the analogy between the Cauchy integral over $|z| = \text{const}$ with measure $z^{-j-1} dz$ and the Laplace inversion integral for the time $t = t_j = jk$ over $\text{Im } \omega = \text{const}$ with measure $e^{-i\eta\omega} d\omega$.

This operational quadrature method can be applied at several different stages of an integral equation method for the time-discretized initial value problem.

It can be used to find a fundamental solution for the whole system in the form of a Cauchy integral over the frequency-domain fundamental solutions G_ω . We get for the coefficients G_j of the semidiscrete space-time fundamental solution $\mathcal{G}(i\partial_t^k)$ the formula

$$G_j(x) = \frac{1}{2\pi i} \int_{|z|=\rho} G_{\omega(z)}(x) z^{-j-1} dz \quad \text{with} \quad \omega(z) = i\frac{\delta(z)}{k}$$

This integral over holomorphic functions can be evaluated numerically with high speed and high accuracy using the trapezoidal rule and FFT. In simple cases, it can be evaluated analytically, for example, in the case of the backward Euler method, where we have the simple characteristic function

$$\delta(z) = 1 - z$$

The Cauchy integral then gives the higher-order fundamental solutions $G^{(j)}$ of the previous section.

This fundamental solution $\mathcal{G}(i\partial_t^k)$ can then be used in a standard boundary element method, keeping in mind that the time-discretized solution will be obtained by finite convolution.

The operational quadrature scheme can also (and equivalently) be introduced at a later stage in the integral equation method, after the frequency-domain integral equations have been solved. Let us describe this with the example of the single-layer representation method for the initial-Dirichlet problem of the heat equation.

The space-time single-layer heat potential operator on Σ can be written as $V = \widehat{V}(i\partial_t)$, where $\widehat{V}(\omega)$ is the frequency-domain single-layer potential operator on Γ whose kernel is the fundamental solution of the Helmholtz operator $(-\Delta - \omega^2)$. Inverting V amounts to evaluating the Cauchy integral of the inverse z -transform where the frequency-domain single-layer integral equations have been solved for those frequencies needed for the Cauchy integral. For the approximation ψ_n of the solution $\psi(t_n)$ at the time $t_n = nk$

with time step k and a space discretization $V_h(\omega)$ of $V(\omega)$, one then obtains

$$\psi_n = \frac{1}{2\pi} \int_{|z|=\rho} V_h\left(i\frac{\delta(z)}{k}\right)^{-1} \left(\sum_{j=0}^n s_{n-j} z^{-j-1} \right) dz \quad (17)$$

This can be compared to the Laplace inversion integral (5) where the contour C is the image of the circle $|z| = \rho$ under the mapping $z \mapsto \omega = i(\delta(z)/k)$. When the Cauchy integral in (17) is evaluated numerically by a quadrature formula, we obtain an end result that has a form very similar to what we got from the Laplace transform boundary element method in formula (7).

In the papers, Lubich and Schneider (1992) and Lubich (1994), the operational quadrature method has been analyzed for a large class of parabolic and hyperbolic initial-boundary value problems and multistep methods satisfying various stability conditions. Recent computational results show its efficiency in practice (Schanz and Antes, 1997a,b).

REFERENCES

- Abate J and Whitt W. Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Comput.* 1995; 7:36–43.
- Adolfsson V, Jawerth B and Torres R. A boundary integral method for parabolic equations in non-smooth domains. *Commun. Pure Appl. Math.* 1994; 47(6):861–892.
- Aliahi MH and Wrobel LC. *The Boundary Element Method*. John Wiley & Sons: New York, 2002.
- Antes H. A boundary element procedure for transient wave propagations in two-dimensional isotropic elastic media. *Finite Elem. Anal. Des.* 1985; 1:313–322.
- Antes H. *Anwendungen der Methode der Randlelemente in der Elastodynamik*. Teubner: Stuttgart, 1988.
- Arendt W, Batty C, Hieber M and Neubrander F. *Vector-Valued Laplace Transforms and Cauchy Problems*. Birkhäuser Verlag: Basel, 2001.
- Arnold DN and Noon PJ. Coercivity of the single layer heat potential. *J. Comput. Math.* 1989; 7:100–104.
- Bachelot A and Lange V. Time dependent integral method for Maxwell's system. *Mathematical and Numerical Aspects of Wave Propagation (Mandelieu-La Napoule, 1995)*. SIAM: Philadelphia, 1995; 151–159.
- Bachelot A, Bounhoure L and Pujols A. Couplage éléments finis-potentiels retardés pour la diffraction électromagnétique par un obstacle hétérogène. *Numer. Math.* 2001; 89(2):257–306.
- Bamberger A and Ha Duong T. Formulation variationnelle espace-temps pour le calcul par potentiel retardé d'une onde acoustique. *Math. Methods Appl. Sci.* 1986; 8:405–435, 598–608.
- Becache E. Résolution par une méthode d'équations intégrales d'un problème de diffraction d'ondes élastiques transitoires par une fissure. Thèse de doctorat, Université Paris VI, 1991.
- Becache E. A variational boundary integral equation method for an elastodynamic antiplane crack. *Int. J. Numer. Methods Eng.* 1993; 36(6):969–984.
- Becache E and Ha-Duong T. A space-time variational formulation for the boundary integral equation in a 2D elastic crack problem. *RAIRO Modél. Math. Anal. Numér.* 1994; 28(2):141–176.
- Birgisson B, Siebrits E and Petroc AP. Elastodynamic direct boundary element methods with enhanced numerical stability properties. *Int. J. Numer. Methods Eng.* 1999; 46(6):871–888.
- Brebbia CA, Telles JCF and Wrobel LC. *Boundary Element Techniques*. Springer-Verlag: Berlin, 1984.
- Brown RM. The method of layer potentials for the heat equation in Lipschitz cylinders. *Am. J. Math.* 1989; 111:339–379.
- Brown RM and Shen ZW. A note on boundary value problems for the heat equation in Lipschitz cylinders. *Proc. Am. Math. Soc.* 1993; 119(2):585–594.
- Bruno O. Fast, high-order, high-frequency integral methods for computational acoustics and electromagnetics. In *Topics in Computational Wave Propagation: Direct and Inverse Problems*, Ainsworth M, Davies P, Duncan D, Martin P, Rynne B (eds). Springer-Verlag: Berlin, 2003; 43–82.
- Chapko R and Kress R. Rothe's method for the heat equation and boundary integral equations. *J. Integral Equations Appl.* 1997; 9(1):47–69.
- Chudinovich I. The solvability of boundary equations in mixed problems for non-stationary Maxwell system. *Math. Methods Appl. Sci.* 1997; 20(5):425–448.
- Chudinovich I. Boundary equations in dynamic problems of the theory of elasticity. *Acta Appl. Math.* 2001; 65(1–3):169–183. Special issue dedicated to Antonio Avantaggiato on the occasion of his 70th birthday.
- Chudinovich IY. The boundary equation method in the third initial-boundary value problem of the theory of elasticity. I. Existence theorems. *Math. Methods Appl. Sci.* 1993a; 16(3):203–215.
- Chudinovich IY. The boundary equation method in the third initial-boundary value problem of the theory of elasticity. II. Methods for approximate solutions. *Math. Methods Appl. Sci.* 1993b; 16(3):217–227.
- Chudinovich IY. On the solution by the Galerkin method of boundary equations in problems of nonstationary diffraction of elastic waves by three-dimensional cracks. *Differentsial'nye Uravneniya* 1993c; 29(9):1648–1651, 1656.
- Costabel M. Boundary integral operators on Lipschitz domains: Elementary results. *SIAM J. Math. Anal.* 1988; 19:613–626.
- Costabel M. Boundary integral operators for the heat equation. *Integral Equations Operator Theory* 1990; 13:498–522.
- Costabel M and Dauge M. On representation formulas and radiation conditions. *Math. Methods Appl. Sci.* 1997; 20:133–150.
- Costabel M and Saranen J. Spline collocation for convolutional parabolic boundary integral equations. *Numer. Math.* 2000; 84(3):417–449.

- Costabel M and Saranen J. Parabolic boundary integral operators: symbolic representation and basic properties. *Integral Equations Operator Theory* 2001; **40**(2):185–211.
- Costabel M and Saranen J. The spline collocation method for parabolic boundary integral equations on smooth curves. *Numer. Math.* 2003; **93**(3):549–562.
- Costabel M, Onishi K and Wendland WL. A boundary element collocation method for the Neumann problem of the heat equation. In *Inverse and Ill-Posed Problems*, Engl HW, Groetsch CW (eds). Academic Press: Boston, 1987; 369–384.
- Cruse T. A direct formulation and numerical solution of the general transient elastodynamic problem. II. *J. Math. Anal. Appl.* 1968; **22**:441–455.
- Cruse T and Rizzo F. A direct formulation and numerical solution of the general transient elastodynamic problem I. *J. Math. Anal. Appl.* 1968; **22**:444–259.
- Dahlberg B and Verchota G. Galerkin methods for the boundary integral equations of elliptic equations in non-smooth domains. *Contemp. Math.* 1990; **107**:39–60.
- Däschle C. Eine Raum-Zeit-Variationsformulierung zur Bestimmung der Potentiale für das elektromagnetische Streuproblem im Außenraum. Dissertation, Universität Freiburg, 1992.
- Davies PJ. Numerical stability and convergence of approximations of retarded potential integral equations. *SIAM J. Numer. Anal.* 1994; **31**(3):856–875.
- Davies PJ. A stability analysis of a time marching scheme for the general surface electric field integral equation. *Appl. Numer. Math.* 1998; **27**(1):33–57.
- Davies PJ and Duncan DB. Averaging techniques for time-marching schemes for retarded potential integral equations. *Appl. Numer. Math.* 1997; **23**(3):291–310.
- Davies PJ and Duncan DB. Stability and convergence of collocation schemes for retarded potential integral equations. Preprint NI03020. Newton Institute: Cambridge, 2003; *SIAM J. Numer. Anal.* to appear.
- Davies B and Martin B. Numerical inversion of the Laplace transform: a survey and comparison of methods. *J. Comput. Phys.* 1979; **33**(1):1–32.
- Faul AC and Powell MJD. Proof of convergence of an iterative technique for thin plate spline interpolation in two dimensions. *Adv. Comput. Math.* 1999; **11**(2–3):183–192. Radial basis functions and their applications.
- Gaul L and Schanz M. A comparative study of three boundary element approaches to calculate the transient response of viscoelastic solids with unbounded domains. *Comput. Methods Appl. Mech. Eng.* 1999; **179**(1–2):111–123.
- Gaut L, Kögl M and Wagner M. *Boundary Element Methods for Engineers and Scientists*. Springer-Verlag: Berlin, 2003.
- Greengard L and Lin P. Spectral approximation of the free-space heat kernel. *Appl. Comput. Harmon. Anal.* 2000; **9**(1):83–97.
- Greengard L and Strain J. A fast algorithm for the evaluation of heat potentials. *Commun. Pure Appl. Math.* 1990; **43**(8):949–963.
- Ha-Duong T. On the transient acoustic scattering by a flat object. *Jpn. J. Appl. Math.* 1990; **7**(3):489–513.
- Ha-Duong T. On boundary integral equations associated to scattering problems of transient waves. *Z. Angew. Math. Mech.* 1996; **76**(Suppl. 2):261–264.
- Ha-Duong T. On retarded potential boundary integral equations and their discretisation. In *Topics in Computational Wave Propagation: Direct and Inverse Problems*, Ainsworth M, Davies P, Duncan D, Martin P, Rynne B (eds). Springer-Verlag: Berlin, 2003; 301–336.
- Hamina M and Saranen J. On the spline collocation method for the single-layer heat operator equation. *Math. Comp.* 1994; **62**(2):41–64.
- Hebekker F-K and Hsiao GC. On Volterra boundary integral equations of the first kind for nonstationary Stokes equations. *Advances in boundary element techniques*, Springer Ser. Comput. Mech. Springer: Berlin, 1993; 173–186.
- Hsiao GC and Saranen J. Boundary integral solution of the two-dimensional heat equation. *Math. Methods Appl. Sci.* 1993; **16**(2):87–114.
- Jiao D, Ergin AA, Balasubramaniam S, Michielssen E and Jin J-M. A fast higher-order time-domain finite element-boundary integral method for 3-D electromagnetic scattering analysis. *IEEE Trans. Antennas Propagation* 2002; **50**(9):1192–1202.
- Khatouryansky NM and Sosa H. Dynamic representation formulas and fundamental solutions for piezoelectricity. *Int. J. Solids Struct.* 1995; **32**(22):3307–3325.
- Kress R. *Linear Integral Equations*. Springer-Verlag: Berlin, 1989.
- Kupradze VD, Gegelia TG, Basheleishvili MO and Burchuladze TV. *Three-Dimensional Problems of the Mathematical Theory of Elasticity And Thermoelasticity*, Vol. 25 of North-Holland Series in Applied Mathematics and Mechanics (Russian edn). North Holland: Amsterdam, 1979.
- Lubich C. On the multistep time discretization of linear initial-boundary value problems and their boundary integral equations. *Numer. Math.* 1994; **67**:365–390.
- Lubich C and Schneider R. Time discretisation of parabolic boundary integral equations. *Numer. Math.* 1992; **63**:455–481.
- Manusur WJ. *A Time-Stepping Technique to Solve Wave Propagation Problems Using the Boundary Element Method*. PhD thesis, University of Southampton, 1983.
- Michielssen E. Fast evaluation of three-dimensional transient wave fields using diagonal translation operators. *J. Comput. Phys.* 1998; **146**(1):157–180.
- Michielssen E, Ergin A, Shanker B and Weile D. The multilevel plane wave time domain algorithm and its applications to the rapid solution of electromagnetic scattering problems: a review. *Mathematical and Numerical Aspects of Wave Propagation (Santiago de Compostela, 2000)*. SIAM: Philadelphia, 2000; 24–33.
- Nardini D and Brebbia CA. A new approach to free vibration analysis using boundary elements. *Appl. Math. Modell.* 1983; **7**:157–162.
- Partidge PW, Brebbia CA and Wrobel LC. *The Dual Reciprocity Boundary Element Method*, International Series on Computational Engineering, Computational Mechanics Publications: Southampton, 1992.
- Peirce A and Siebrits E. Stability analysis of model problems for elastodynamic boundary element discretizations. *Numer. Methods Partial Differential Equations* 1996; **12**(5):585–613.

- Peirce A and Siebrits E. Stability analysis and design of time-stepping schemes for general elastodynamic boundary element models. *Int. J. Numer. Methods Eng.* 1997; **40**(2):319–342.
- Pogorzelski W. *Integral Equations and their Applications*. Pergamon Press: Oxford, 1966.
- Powell MJD. The theory of radial basis function approximation in 1990. *Advances in Numerical Analysis*, vol. II (Lancaster, 1990). Oxford Science Publications, Oxford University Press: New York, 1992; 105–210.
- Pujols A. *Equations intégrales espace-temps pour le système de Maxwell – application au calcul de la surface équivalente Radar*. Thèse de doctorat, Université Bordeaux I, 1991.
- Rynne BP. The well-posedness of the electric field integral equation for transient scattering from a perfectly conducting body. *Math. Methods Appl. Sci.* 1999; **22**(7):619–631.
- Sayah T. *Méthodes de potentiels retardés pour les milieux hétérogènes et l'approximation des couches minces par conditions d'impédance généralisées en électromagnétisme*. Thèse de doctorat, Ecole Polytechnique, 1998.
- Schanz M. A boundary element formulation in time domain for viscoelastic solids. *Commun. Numer. Methods Eng.* 1999; **15**(11):799–809.
- Schanz M. Application of 3D time domain boundary element formulation to wave propagation in poroelastic solids. *Eng. Anal. Bound. Elem.* 2001a; **25**(4–5):363–376.
- Schanz M. *Wave Propagation in Viscoelastic and Poroelastic Continua: A Boundary Element Approach*, Lecture Notes in Applied and Computational Mechanics, Springer-Verlag: Berlin-Heidelberg-New York, 2001b.
- Schanz M and Antes H. Application of 'operational quadrature methods' in time domain boundary element methods. *Meccanica* 1997a; **32**(3):179–186.
- Schanz M and Antes H. A new visco- and elastodynamic time domain Boundary Element formulation. *Comput. Mech.* 1997b; **20**(5):452–459.
- Telles JCF and Vera-Tudela CAR. A BEM NQF technique coupled with the operational quadrature method to solve elastodynamic crack problems. In *Advances in Boundary Element Techniques IV*, Gallego R, Aliabadi MH (eds). Department of Engineering, Queen Mary, University of London: London, 2003; 1–6.
- Terrasse I. *Résolution mathématique et numérique des équations de Maxwell instationnaires par une méthode de potentiels retardés*. Thèse de doctorat, Ecole Polytechnique, 1993.
- Wang CY, Zhang C and Hirose S. Dynamic fundamental solutions and time-domain BIE formulations for piezoelectric solids. In *Advances in Boundary Element Techniques IV*, Gallego R, Aliabadi MH (eds). Department of Engineering, Queen Mary, University of London: London, 2003; 215–224.

Chapter 26

Finite Element Methods for Maxwell Equations

Leszek Demkowicz
The University of Texas at Austin, Austin, TX, USA

| | |
|---|-----|
| 1 Maxwell Equations | 723 |
| 2 Variational Formulation | 725 |
| 3 Exact Sequences | 727 |
| 4 Projection-based Interpolation. De Rham Diagram | 732 |
| 5 Additional Comments | 734 |
| 6 Related Chapters | 735 |
| Acknowledgment | 735 |
| Notes | 735 |
| References | 736 |
| Further Reading | 737 |

1 MAXWELL EQUATIONS

We shall discuss the simplest (linear, isotropic) version of Maxwell's equations. Given a domain $\Omega \subset \mathbb{R}^3$, we wish to determine electric field $E(x)$ and magnetic field $H(x)$ that satisfy:

- Faraday's law (1831),

$$\nabla \times E = -\frac{\partial}{\partial t}(\mu H)$$

- Ampere's law (1820) with Maxwell's correction (1856),

$$\nabla \times H = J^{imp} + \sigma E + \frac{\partial}{\partial t}(\epsilon E)$$

Encyclopedia of Computational Mechanics, Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes. Volume 1: *Fundamentals*. © 2004 John Wiley & Sons, Ltd. ISBN: 0-470-84699-2.

Here μ, σ, ϵ denote the material data: permeability, conductivity, and permittivity, assumed to be piecewise constant. $B := \mu H$ is the magnetic flux, $D := \epsilon E$ is the electric flux, σE is the Ohm current, and J^{imp} denotes a prescribed, given impressed current, with $J := J^{imp} + \sigma E$ identified as the total current.

Once the total current J has been determined, we can use the continuity (conservation of free charge) equation to determine the corresponding free charge density ρ ,

$$\nabla \cdot J + \frac{\partial \rho}{\partial t} = 0$$

The first-order Maxwell system is accompanied with initial, boundary, and interface (across material discontinuities) conditions.

Taking the divergence of both sides in Faraday's equation, we learn that the magnetic field H has to satisfy (automatically) the equation,

$$\frac{\partial}{\partial t}(\nabla \cdot (\mu H)) = 0$$

Assuming that the initial value $H(0)$ satisfies the Gauss law for the magnetic flux,

$$\nabla \cdot (\mu H) = 0$$

we learn that the law is satisfied automatically for all times t .

Similarly, taking the divergence of both sides of the Ampere equation, and utilizing the continuity law, we learn that the electric field E satisfies,

$$\frac{\partial}{\partial t}(\nabla \cdot (\epsilon E) - \rho) = 0$$

Assuming that the initial value $E(0)$ satisfies the Gauss law for the electric flux,

$$\nabla \cdot (\epsilon E) = \rho$$

we learn that the electric field satisfies the law at all times t . In the steady state, the Maxwell system degenerates and decouples,

$$\begin{cases} \nabla \times E = 0 \\ \nabla \cdot (\epsilon E) = \rho \end{cases}$$

The closing equation for electrostatics is provided either by the continuity equation or by the Gauss law. In the case of a perfect dielectric, $\sigma = 0$, the distribution of free charge ρ must be prescribed. We can determine the corresponding electric field by solving the electrostatics equations,

$$\begin{cases} \nabla \times E = 0 \\ \nabla \cdot (\epsilon E) = \rho \end{cases}$$

In the case of a conductor, $\sigma > 0$, the free charges move, and we cannot prescribe them. In the steady state case, ρ is independent of time, and the continuity equation provides a closing equation for the Faraday law,

$$\begin{cases} \nabla \times E = 0 \\ -\nabla \cdot (\sigma E) = \nabla \cdot J^{imp} \end{cases}$$

Once the electric field is known, the resulting free charge density can be determined from the Gauss law. In view of the first equation, either set of the electrostatics equations is usually solved in terms of a scalar potential for the electric field.

In the case of a perfect conductor, $\sigma = \infty$, the corresponding electric field E vanishes, and the volume occupied by the perfect conductor is eliminated from the (computational) domain Ω .

The magnetostatics equations can be obtained by complementing the Ampere law with the Gaussian law for the magnetic flux.

$$\begin{cases} \nabla \times H = J^{imp} + \sigma E \\ \nabla \cdot (\mu H) = 0 \end{cases}$$

Once the electric field is known, the corresponding magnetic field is obtained by solving the magnetostatics equations. Because of the second equation, the problem is usually formulated in terms of a vector potential for the magnetic flux $B = \mu H$.

Wave equation

In order to reduce the number of unknowns, the first-order Maxwell system is usually reduced to a single (vector-valued) 'wave equation', expressed either in terms of E

or H . The choice is usually dictated by the boundary conditions, and the analysis of both systems is fully analogous. We shall focus on the electric field formulation,

$$\nabla \times \left(\frac{1}{\mu} \nabla \times E \right) + \epsilon \frac{\partial^2 E}{\partial t^2} + \sigma \frac{\partial E}{\partial t} = -\frac{\partial J^{imp}}{\partial t} \quad (1)$$

Once the electric field has been determined, the Faraday equation can be integrated to find the corresponding magnetic field.

Time-harmonic wave equation

Assuming the ansatz,

$$E(x, t) = \Re(E(x)e^{i\omega t}) \quad (2)$$

we convert the wave equation into the 'reduced wave equation',

$$\nabla \times \left(\frac{1}{\mu} \nabla \times E \right) - (\omega^2 \epsilon - i\omega \sigma) E = -i\omega J^{imp} \quad (3)$$

to be solved for the complex-valued phasor $E(x)$. Alternatively, (3) can be obtained by applying Fourier transform to (1). The solution to the wave equation can then be obtained by applying the inverse Fourier transform,

$$E(x, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t} E(x, \omega) d\omega$$

Notice that the sign in the exponential in the inverse Fourier transform is consistent with that in the ansatz (2). In the Electrical Engineering (EE) literature, frequently, the opposite sign is assumed,

$$E(x, t) = \Re(E(x)e^{-i\omega t}) \quad (4)$$

with j denoting the imaginary unit. The sign in the ansatz affects the sign in impedance and radiation boundary conditions, and one has to always remember which ansatz is being used. Substituting $j = -i$, we can easily switch in between the two formulations.

Once the electric field has been determined, the corresponding magnetic field is computed from the time-harmonic version of the Faraday law:

$$\nabla \times E = -i\omega \mu H \quad (5)$$

For free space, $\epsilon = \epsilon_0 \approx (1/36\pi)10^{-9}$ [C²/Nm² = F/m], $\mu = \mu_0 = 4\pi 10^{-7}$ [N/A² = h/m], $\sigma = 0$, and (3) reduces to:

$$\nabla \times (\nabla \times E) - \left(\frac{\omega}{c} \right)^2 E = -i\omega \mu_0 J^{imp}$$

Table 1. Material constants for selected materials.

| Material | μ_r | ϵ_r | $\sqrt{\mu_r/\epsilon_0}$ |
|------------------------|---------|--------------|---------------------------|
| Alumina (10 GHz) | 1 | 10 | 0.63 |
| Polyethylene (10 GHz) | 1 | 2.25 | 0.19 |
| Copper | 1 | 1 | 2.2E+9 |
| Seawater | 1 | 1 | 1.5E+3 |
| Human muscle (900 MHz) | 1 | 58 | 4.56E+2 |

where $c = (\epsilon_0 \mu_0)^{1/2} = 3 \times 10^8$ ms⁻¹ is the speed of light in free space and $k_0 = \omega/c$ (1/m) is the free space wave number. Introducing relative permittivity, and relative permeability,

$$\epsilon_r = \frac{\epsilon}{\epsilon_0}, \quad \mu_r = \frac{\mu}{\mu_0}$$

we represent the general case in the form,

$$\begin{aligned} \nabla \times \left(\frac{1}{\mu_r} \nabla \times E \right) - \left(k_0^2 \epsilon_r - i k_0 \sqrt{\frac{\mu_0}{\epsilon_0}} \sigma \right) E \\ = -i k_0 \sqrt{\frac{\mu_0}{\epsilon_0}} J^{imp} \end{aligned} \quad (6)$$

In reality, the material constants depend upon frequency ω , temperature, and other factors. The discussed equations apply only to isotropic materials, many materials, for example, ferromagnetics, are strongly anisotropic. A few sample values of relative permeability μ_r , relative permittivity ϵ_r , and scaled conductivity $\sqrt{(\mu_0/\epsilon_0)}\sigma$ are summarized in Table 1.

We shall now return to our original notation (3), with the understanding that, for practical computations, we use formulation (6).

2 VARIATIONAL FORMULATION

For the sake of simplicity, we shall restrict our presentation to the case of bounded, simply connected domains Ω only. We shall focus on the time-harmonic Maxwell equations. We begin with the fundamental 'integration by parts' formula,

$$\int_{\Omega} (\nabla \times E) F dx = \int_{\Omega} E (\nabla \times F) dx + \int_{\partial\Omega} (n \times E) F_t dS$$

Here n is the outward normal unit vector for boundary $\partial\Omega$, $F_t = F - (F \cdot n)n$ is the tangential component of vector F on the boundary, and

$$n \times E = n \times E_t$$

is the 'rotated' tangential component of E . Obviously, $(n \times E) F_t = E_t (n \times F)$. Notice that there is no '-' sign typical of other Green's formulas.

Finite energy considerations lead us naturally to the assumption that both electric field and magnetic field are square integrable. In view of (5), and under the assumption of boundedness of material data,

$$0 < \epsilon_{\min} \leq \epsilon \leq \epsilon_{\max} < \infty, \quad 0 < \mu_{\min} \leq \mu \leq \mu_{\max} < \infty \\ 0 \leq \sigma \leq \sigma_{\max} < \infty$$

this implies that electric field E comes from the $H(\text{curl})$ space,

$$H(\text{curl}, \Omega) = \{E \in L^2(\Omega); \nabla \times E \in L^2(\Omega)\}$$

The Green formula is immediately telling us the right type of continuity across material interfaces and interelement boundaries for a conforming finite element (FE) discretization. Assuming that domain Ω consists of two disjoint parts Ω_1, Ω_2 , with an interface Γ , with a C^1 field E in either of the subdomains, we use the integration by parts formula to obtain,

$$\int_{\Omega} (\nabla \times E) \phi dx = \int_{\Omega} E (\nabla \times \phi) dx + \int_{\Gamma} [n \times E] \phi dS$$

for every C^1 test function ϕ vanishing on $\partial\Omega$. Here $[n \times E]$ denotes the jump of the tangential component of E across the interface Γ . Consequently, the field $\nabla \times E$ is a function (regular distribution) if and only if the tangential component of E is continuous across the interface. With a square integrable impressed current J^{imp} , similar considerations for the magnetic field lead to the observation that also the tangential component of H must be continuous across the material interfaces. In view of (5), this implies the second interface condition for the electric field,

$$\left[n \times \frac{1}{\mu} (\nabla \times E) \right] = 0$$

Multiplying (3) with a (conjugated) test function F , and integrating by parts, we obtain,

$$\begin{aligned} \int_{\Omega} \left\{ \frac{1}{\mu} (\nabla \times E) (\nabla \times \bar{F}) - (\omega^2 \epsilon - i\omega \sigma) E \bar{F} \right\} dx \\ + \int_{\partial\Omega} n \times \frac{1}{\mu} (\nabla \times E) \bar{F}_t dS = -i\omega \int_{\Omega} J^{imp} \bar{F} dx \end{aligned} \quad (7)$$

Working with conjugated test functions that result in sesquilinear rather than bilinear forms in the variational formulation, is typical for complex-valued wave propagation

problems, and consistent with the definition of inner product in a complex Hilbert space. It is only a matter of choice though and, as long as shape functions are real-valued, the bilinear and sesquilinear formulations yield identical systems of discrete equations.

We are now ready to discuss the most common boundary conditions.

Perfect electric conductor (PEC)

As the electric field in a perfect conductor vanishes, and the tangential component of E must be continuous across material interfaces, the tangential component of E on a boundary adjacent to a perfect conductor must vanish,

$$n \times E = 0$$

For scattering problems, the electric field is the sum of a known incident field E^{inc} and (to be determined) scattered field E^s . The condition above then leads to a nonhomogeneous Dirichlet condition for the scattered field,

$$n \times E^s = -n \times E^{\text{inc}}$$

Impressed surface current

The simplest way to model an antenna is to prescribe an impressed surface current on the boundary of the domain occupied by the antenna,

$$n \times H = J_s^{\text{imp}}$$

In conjunction with (5), this leads to the Neumann (natural boundary) condition,

$$n \times \frac{1}{\mu} (\nabla \times E) = -i\omega J_s^{\text{imp}}$$

For a particular case of $J_s^{\text{imp}} = 0$, we speak of a magnetic symmetry wall condition.

Impedance boundary condition

This is the simplest, first-order approximation to model reflection of the electric field from an interface with a conductor with large but nevertheless finite conductivity, comp. Senior and Volakis (1995),

$$n \times \frac{1}{\mu} (\nabla \times E) - i\omega \gamma E_t = -i\omega J_s^{\text{imp}}$$

with a constant $\gamma > 0$. One arrives naturally at the impedance boundary condition also when modeling waveguides.

Denoting by $\Gamma_1, \Gamma_2, \Gamma_3$, three disjoint components of boundary $\partial\Omega$, on which the Dirichlet, Neumann, and

impedance boundary conditions have been prescribed, we limit the test functions in (7) to those that satisfy the homogeneous PEC condition on Γ_1 , and use Neumann and Cauchy conditions to build the impressed surface currents into the formulation. With appropriate regularity assumptions on the impressed currents, for example, both J_s^{imp} and J_s^{imp} being square integrable, our final variational formulation reads as follows.

$$\begin{cases} E \in H(\text{curl}, \Omega), n \times E = n \times E_0 \text{ on } \Gamma_1 \\ \int_{\Omega} \left\{ \frac{1}{\mu} (\nabla \times E)(\nabla \times \bar{F}) - (\omega^2 \epsilon - i\omega\sigma) E \cdot \bar{F} \right\} dx \\ + i\omega \int_{\Gamma_2} \gamma E_t \bar{F} dS = -i\omega \int_{\Omega} J_s^{\text{imp}} \bar{F} dx \\ + i\omega \int_{\Gamma_3 \cup \Gamma_1} J_s^{\text{imp}} \bar{F} dS \quad \text{for every } F \in H(\text{curl}, \Omega), \\ n \times F = 0 \text{ on } \Gamma_1 \end{cases} \quad (8)$$

We follow the classical lines to show that, conversely, any sufficiently regular solution to the variational problem satisfies the reduced wave equation and the natural boundary conditions.

Weak form of the continuity equation

Employing a special test function, $F = \nabla q, q \in H^1(\Omega)$, $q = 0$ on Γ_1 , we learn that the solution to the variational problem automatically satisfies the weak form of the continuity equation,

$$\begin{aligned} \int_{\Omega} -(\omega^2 \epsilon - i\omega\sigma) E \nabla \bar{q} dx + i\omega \int_{\Gamma_2} \gamma E_t \nabla \bar{q} dS \\ = -i\omega \int_{\Omega} J_s^{\text{imp}} \nabla \bar{q} dx + i\omega \int_{\Gamma_3 \cup \Gamma_1} J_s^{\text{imp}} \nabla \bar{q} dS \\ \text{for every } q \in H^1(\Omega), q = 0 \text{ on } \Gamma_1 \end{aligned} \quad (9)$$

Upon integrating by parts, we learn that solution E satisfies the continuity equation,

$$\text{div}((\omega^2 \epsilon - i\omega\sigma) E) = i\omega \text{div } J^{\text{imp}} \quad (= \omega^2 \rho)$$

plus additional boundary conditions on Γ_2, Γ_3 , and interface conditions across material interfaces.

Maxwell eigenvalue problem

Related to the time-harmonic problem (8) is the eigenvalue problem,

$$\begin{cases} E \in H(\text{curl}, \Omega), n \times E = 0 \text{ on } \Gamma_1, \quad \lambda \in \mathbb{R} \\ \int_{\Omega} \frac{1}{\mu} (\nabla \times E)(\nabla \times \bar{F}) dx = \lambda \int_{\Omega} \epsilon E \cdot \bar{F} dx \\ \text{for every } F \in H(\text{curl}, \Omega), n \times F = 0 \text{ on } \Gamma_1 \end{cases} \quad (10)$$

The curl-curl operator is self-adjoint. Its spectrum consists of $\lambda = 0$ with an infinite-dimensional eigenspace consisting of all gradients $\nabla p, p \in H^1(\Omega), p = 0$ on Γ_1 , and a sequence of positive eigenvalues $\lambda_1 < \lambda_2 < \dots < \lambda_n \rightarrow \infty$ with corresponding eigenspaces of finite dimension. Only the eigenvectors corresponding to positive eigenvalues are physical. Repeating the reasoning with the substitution $F = \nabla q$, we learn that they automatically satisfy the continuity equation.

Stabilized variational formulation

The standard variational formulation (8) is not uniformly stable with respect to frequency ω . As $\omega \rightarrow 0$, we loose control over gradients. This corresponds to the fact that, in the limiting case $\omega = 0$, the problem is ill-posed as the gradient component remains undetermined. A remedy to this problem is to enforce the continuity equation explicitly at the expense of introducing a Lagrange multiplier p . The so-called stabilized variational formulation looks as follows.

$$\begin{cases} E \in H(\text{curl}, \Omega), p \in H^1(\Omega), \quad n \times E = n \times E_0 \\ p = 0 \text{ on } \Gamma_1 \\ \int_{\Omega} \frac{1}{\mu} (\nabla \times E)(\nabla \times \bar{F}) dx - \int_{\Omega} (\omega^2 \epsilon - i\omega\sigma) E \cdot \bar{F} dx \\ + i\omega \int_{\Gamma_2} \gamma E_t \bar{F} dS - \int_{\Omega} (\omega^2 \epsilon - i\omega\sigma) \nabla p \cdot \bar{F} dx \\ = -i\omega \int_{\Omega} J_s^{\text{imp}} \cdot \bar{F} dx + i\omega \int_{\Gamma_3 \cup \Gamma_1} J_s^{\text{imp}} \cdot \bar{F} dS \\ \forall F \in H(\text{curl}, \Omega), n \times F = 0 \text{ on } \Gamma_1 \\ - \int_{\Omega} (\omega^2 \epsilon - i\omega\sigma) E \cdot \nabla \bar{q} dx + i\omega \int_{\Gamma_2} \gamma E_t \nabla \bar{q} dS = \\ -i\omega \int_{\Omega} J_s^{\text{imp}} \cdot \nabla \bar{q} dx + i\omega \int_{\Gamma_3 \cup \Gamma_1} J_s^{\text{imp}} \cdot \nabla \bar{q} dS \\ \forall q \in H^1(\Omega), q = 0 \text{ on } \Gamma_1 \end{cases} \quad (11)$$

By repeating the reasoning with the substitution $F = \nabla q$ in the first equation, we learn that the Lagrange multiplier p satisfies the weak form of a Laplace-like equation with homogeneous boundary conditions and, therefore, it *identically vanishes*. For this reason, it is frequently called the *hidden variable*. The stabilized formulation has improved stability properties for small ω . In the case of $\sigma = 0$ and right-hand side of (9) vanishing, we can rescale the Lagrange multiplier, $p = \omega^2 \tilde{p}, q = \omega^2 \tilde{q}$, to obtain a symmetric mixed variational formulation with stability constant converging to one as $\omega \rightarrow 0$. In the general case, we cannot avoid a degeneration as $\omega \rightarrow 0$ but we can still rescale the Lagrange multiplier with ω ($p = \omega \tilde{p}, q = \omega \tilde{q}$), to improve the stability of the formulation for small ω . The stabilized formulation is possible because gradients of the scalar-valued potentials from

$H^1(\Omega)$ form precisely the null space of the curl-curl operator.

The point about the stabilized (mixed) formulation is that, whether we use it or not in the actual computations (the improved stability is one good reason to do it...), the original variational problem is *equivalent* to the mixed problem. This suggests that we cannot escape from the theory of mixed formulations when analyzing the problem.

3 EXACT SEQUENCES

The gradient and curl operators, along with the divergence operator, form an *exact sequence*,

$$\mathbb{R} \rightarrow H^1 \xrightarrow{\nabla} H(\text{curl}) \xrightarrow{\nabla \times} H(\text{div}) \xrightarrow{\nabla \cdot} L^2 \rightarrow 0$$

In an exact sequence of operators, the range of each operator coincides with the null space of the operator next in the sequence. Simply speaking, the gradient of a function vanishes if and only if the function is constant, the curl of a vector-valued function is zero, if and only if the function is the gradient of a scalar potential, and so on. The spaces above may incorporate homogeneous essential boundary conditions. Introducing,

$$\begin{aligned} W &= \{q \in H^1(\Omega): q = 0 \text{ on } \Gamma_1\} \\ Q &= \{E \in H(\text{curl}, \Omega), E_t = 0 \text{ on } \Gamma_1\} \\ V &= \{H \in H(\text{div}, \Omega), H_n = 0 \text{ on } \Gamma_1\} \\ Y &= L^2(\Omega) \end{aligned}$$

($H_n = H \cdot n$ denotes normal component of H) we have the exact sequence:

$$W \xrightarrow{\nabla} Q \xrightarrow{\nabla \times} V \xrightarrow{\nabla \cdot} Y \rightarrow 0$$

The presence of \mathbb{R} in the original sequence signifies the fact that the null space of the gradient consists of constant fields. With the Dirichlet boundary condition on boundary Γ_1 , the constant must be zero, and the space of constant fields is eliminated. Similarly, presence of the trivial space at the end of the sequence signifies the fact that the preceding operator is a surjection. If Γ_1 coincides with the *whole boundary*, the L^2 space must be replaced with the space of fields of zero average,

$$L_0^2 = \left\{ u \in L^2(\Omega): \int_{\Omega} u = 0 \right\}$$

In order to simplify the notation, we shall drop \mathbf{R} and the trivial space from our discussion.

In two-space dimensions, the 3D exact sequence gives rise to two sequences,

$$\mathbf{R} \longrightarrow H^1 \xrightarrow{\nabla} H(\text{curl}) \xrightarrow{\text{curl}} L^2$$

and

$$\mathbf{R} \longrightarrow H^1 \xrightarrow{\nabla \times} H(\text{div}) \xrightarrow{\nabla \cdot} L^2$$

Notice the difference between the two curl operators; both are obtained by restricting the 3D curl operator, in the first case, to vectors $(E_1, E_2, 0)$, in the second to $(0, 0, E_3)$, $E_i = E_i(x_1, x_2)$.

$$\text{curl}(E_1, E_2) = E_{1,2} - E_{2,1}, \quad \nabla \times E_3 = (E_{3,2}, -E_{3,1})$$

The second 2D sequence can be obtained from the first one by 'rotating' operators and spaces by 90 degrees.

The exact sequence property is crucial in proving the stability result for the regularized variational formulation for the time-harmonic Maxwell equations; see Demkowicz and Vardapetyan (1998). This suggests constructing (piecewise) polynomial Finite Element discretization of the $H(\text{curl})$ space in such a way that the exact sequence property is also satisfied at the discrete level. Two fundamental families of Nedelec's elements satisfy such a condition.

Tetrahedral element of second type (Nedelec, 1986)

All polynomial spaces are defined on the master tetrahedron,

$$T = \{(x_1, x_2, x_3): x_1 > 0, x_2 > 0, x_3 > 0, \\ x_1 + x_2 + x_3 < 1\}$$

We have the following exact sequence,

$$\mathcal{P}^p \xrightarrow{\nabla} \mathcal{P}^{p-1} \xrightarrow{\nabla \times} \mathcal{P}^{p-2} \xrightarrow{\nabla \cdot} \mathcal{P}^{p-3}$$

Here \mathcal{P}^p denotes the space of polynomials of (group) order less than or equal to p , for example, $x_1^2 x_2^2 x_3^2 \in \mathcal{P}^7$, and $\mathcal{P}^p = \mathcal{P}^p \times \mathcal{P}^p \times \mathcal{P}^p$.

Obviously, the construction starts with $p \geq 3$, that is, the $H(\text{curl})$ -conforming elements are at least of second order.

The construction can be generalized to tetrahedra of variable order. With each tetrahedron's face we associate the corresponding face order p_f , and with each tetrahedron's edge, we associate the corresponding edge order p_e . We assume that,

$$p_f \leq p \quad \forall \text{ face } f, \quad p_e \leq p_f \quad \forall \text{ face } f \text{ adjacent to edge } e, \\ \forall \text{ edge } e$$

The assumption is satisfied in practice by enforcing the *minimum rule*, that is, setting the face and edge orders to the minimum of the orders of the adjacent elements. We introduce now the following polynomial spaces.

- The space of scalar-valued polynomials of order less than or equal to p , whose traces on faces f reduce to polynomials of (possibly smaller) order p_f , and whose traces on edges e reduce to polynomials of (possibly smaller) order p_e ,

$$\mathcal{P}_{p_f, p_e}^p = \{u \in \mathcal{P}^p: u|_f \in \mathcal{P}^{p_f}(f), u|_e \in \mathcal{P}^{p_e}(e)\}$$

- The space of vector-valued polynomials of order less than or equal to p , whose *tangential traces* on faces f reduce to polynomials of order p_f , and whose *tangential traces* on edges e reduce to polynomials of order p_e ,

$$\mathcal{P}_{p_f, p_e}^p = \{E \in \mathcal{P}^p: E|_f \in \mathcal{P}^{p_f}(f), E|_e \in \mathcal{P}^{p_e}(e)\}$$

- The space of vector-valued polynomials of order less than or equal to p , whose *normal traces* on faces f reduce to polynomials of order p_f

$$\mathcal{P}_{p_f}^p = \{E \in \mathcal{P}^p: E_n|_f \in \mathcal{P}^{p_f}(f)\}$$

We have then the exact sequence,

$$\mathcal{P}_{p_f, p_e}^p \xrightarrow{\nabla} \mathcal{P}_{p_f-1, p_e-1}^{p-1} \xrightarrow{\nabla \times} \mathcal{P}_{p_f-2}^{p-2} \xrightarrow{\nabla \cdot} \mathcal{P}^{p-3}$$

with a 2D equivalent for the triangular element,

$$\mathcal{P}_{p_f}^p \xrightarrow{\nabla} \mathcal{P}_{p_f-1}^{p-1} \xrightarrow{\nabla \times} \mathcal{P}^{p-2} \quad (12)$$

The case $p_f, p_e = -1$ corresponds to the homogeneous Dirichlet boundary condition.

Hexahedral element of the first type (Nedelec, 1980)

All polynomial spaces are defined on a unit cube. We introduce the following polynomial spaces.

$$W_p = Q^{(p, q, r)}$$

$$Q_p = Q^{(p-1, q, r)} \times Q^{(p, q-1, r)} \times Q^{(p, q, r-1)}$$

$$V_p = Q^{(p, q-1, r-1)} \times Q^{(p-1, q, r-1)} \times Q^{(p-1, q-1, r)}$$

$$Y_p = Q^{(p-1, q-1, r-1)}$$

Here $Q^{(p, q, r)}$ denotes the space of polynomials of order less than or equal to p, q, r with respect to x, y, z respectively. For instance, $2x^2y^3 + 3x^3z^8 \in Q^{(3, 3, 8)}$. The polynomial spaces form again the exact sequence

$$W_p \xrightarrow{\nabla} Q_p \xrightarrow{\nabla \times} V_p \xrightarrow{\nabla \cdot} Y_p \quad (13)$$

The generalization to variable order elements is a little less straightforward than for the tetrahedra. We shall begin with the 2D case first. For each horizontal edge e , we introduce order p_e , and with each vertical edge e , we associate order q_e . We assume again that the minimum rule holds, that is,

$$p_e \leq p, \quad q_e \leq q$$

By $Q_{p_e, q_e}^{(p, q)}$ we understand the space of polynomials of order less than or equal to p with respect to x , and order less than or equal to q with respect to y , such that their traces to horizontal edges e reduce to polynomials of (possibly smaller) degree p_e , and restrictions to vertical edges reduce to polynomials of (possibly smaller) order q_e .

$$Q_{p_e, q_e}^{(p, q)} = \{u \in Q^{(p, q)}: u(\cdot, 0) \in \mathcal{P}^{p_e}(0, 1), u(\cdot, 1) \in \mathcal{P}^{p_e}(0, 1) \\ u(0, \cdot) \in \mathcal{P}^{q_e}(0, 1), u(1, \cdot) \in \mathcal{P}^{q_e}(0, 1)\}$$

With spaces,

$$W_p = Q_{p_e, q_e}^{(p, q)}$$

$$Q_p = Q_{p_e-1}^{(p-1, q)} \times Q_{q_e-1}^{(p, q-1)}$$

$$Y_p = Q^{(p-1, q-1)}$$

we have the exact sequence,

$$W_p \xrightarrow{\nabla} Q_p \xrightarrow{\nabla \times} Y_p$$

Notice that space Q_p cannot be obtained by merely differentiating polynomials from $Q_{p_e, q_e}^{(p, q)}$. For the derivative in x , this would lead to space $Q_{p_e-1, q_e}^{(p-1, q)}$ for the first component, whereas in our definition above, q_e has been increased to q . This is motivated by the fact that the traces of E_1 along the vertical edges are interpreted as *normal* components of the E field. The $H(\text{curl})$ -conforming fields 'connect' only through tangential components; shape functions corresponding to the normal components on the boundary are classified as interior modes, and they should depend only upon the order of the element and not upon the order of the neighboring elements.

In three dimensions, spaces get more complicated and notation more cumbersome. We start with the space,

$$Q_{p_f, q_f, r_f}^{(p, q, r)}$$

that consists of polynomials in $Q^{(p, q, r)}$ such that:

- their restrictions to faces f parallel to axes x, y reduce to polynomials in $Q^{(p_f, q_f)}$,
- their restrictions to faces f parallel to axes x, z reduce to polynomials in $Q^{(p_f, r_f)}$,

- their restrictions to faces f parallel to axes y, z reduce to polynomials in $Q^{(q_f, r_f)}$,
- their restriction to edges parallel to axis x, y, z reduce to polynomials of order p_e, q_e, r_e respectively,

with the minimum rule restrictions:

$$p_f \leq p, q_f \leq q, r_f \leq r, \quad p_e \leq p_f, q_e \leq q_f, r_e \leq r_f \\ \text{for adjacent faces } f$$

The 3D polynomial spaces forming the de Rham diagram are now introduced as follows:

$$W_p = Q_{p_f, q_f, r_f}^{(p, q, r)}$$

$$Q_p = Q_{p_f-1, q_f}^{(p-1, q, r)} \times Q_{p_f-1, r_f}^{(p-1, q, r)} \times Q_{q_f-1, r_f}^{(p, q-1, r)}$$

$$\times Q_{p_f, q_f-1, r_f}^{(p, q-1, r)}$$

$$\times Q_{p_f, q_f, r_f-1}^{(p, q, r-1)}$$

$$V_p = Q_{q_f-1, r_f-1}^{(p, q-1, r-1)} \times Q_{p_f-1, q_f-1}^{(p-1, q-1, r-1)} \times Q_{p_f-1, q_f-1}^{(p-1, q-1, r)}$$

$$Y_p = Q^{(p-1, q-1, r-1)}$$

Note the following points:

- There is no restriction on edge order in the $H(\text{div})$ -conforming space. The only order restriction is placed on faces normal to the particular component, for example, for the first component H_1 , the order restriction is imposed only on faces parallel to y, z faces.
- For the $H(\text{curl})$ -conforming space, there is no restriction on face order for faces perpendicular to the particular component. For instance, for E_1 , there is no order restriction on faces parallel to y, z axes. The edge orders for edges perpendicular to x are inherited from faces parallel to the x axis. This is related to the fact that elements connecting through the first component E_1 connect only through faces and edges parallel to the first axis.

Tetrahedral element of first type (Nedelec, 1980) There is a significant difference between the tetrahedral and hexahedral elements presented so far. For the tetrahedron, the order p drops in the diagram from p to $p-3$. This merely reflects the fact that the differentiation always lowers the polynomial order by one. In the case of the Q -spaces, however, the order in the diagram has dropped only by one, from (p, q, r) to $(p-1, q-1, r-1)$. A similar effect can be obtained for the tetrahedra.

We shall first discuss the 2D case of a triangular element. The goal is to switch from $p-2$ to $p-1$ in (12) without

increasing the order p in the space on the left. We begin by rewriting (12) with p increased by one.

$$\mathcal{P}_{p_r}^{p+1} \xrightarrow{\nabla} \mathcal{P}_{p_r-1}^p \xrightarrow{\nabla \times} \mathcal{P}^{p-1}$$

Notice that we have not increased the order along the edges. This is motivated by the fact that the edge orders do not affect the very last space in the diagram. Next, we decompose the space of potentials into the previous space of polynomials $\mathcal{P}_{p_r}^p$ and an algebraic complement $\tilde{\mathcal{P}}_{p_r}^{p+1}$,

$$\mathcal{P}_{p_r}^{p+1} = \mathcal{P}_{p_r}^p \oplus \tilde{\mathcal{P}}_{p_r}^{p+1}$$

The algebraic complement is *not unique*, it may be constructed in (infinitely) many different ways. The decomposition in the space of potentials implies a corresponding decomposition in the $H(\text{curl})$ -conforming space,

$$\mathcal{P}_{p_r}^p = \mathcal{P}_{p_r-1}^{p-1} \oplus \nabla(\tilde{\mathcal{P}}_{p_r-1}^{p+1}) \oplus \tilde{\mathcal{P}}_{p_r-1}^p$$

The algebraic complement $\tilde{\mathcal{P}}_{p_r-1}^p$ is again *not unique*. The desired extension of the original sequence can now be constructed by removing the gradients of order $p+1$,

$$\mathcal{P}_{p_r}^p \xrightarrow{\nabla} \mathcal{P}_{p_r-1}^{p-1} \oplus \tilde{\mathcal{P}}_{p_r-1}^p \xrightarrow{\nabla \times} \mathcal{P}^{p-1}$$

Note the following facts.

- The construction enables the $H(\text{curl})$ -conforming discretization of lowest order on triangles. For $p = p_r = 1$,

$$\mathcal{P}_0^0 \oplus \tilde{\mathcal{P}}^1 = \mathcal{P}_0^1, \quad \dim \mathcal{P}_0^1 = 3$$

The complement $\tilde{\mathcal{P}}_1^2$ is empty and, therefore, in this case, the resulting Whitney space $\mathcal{P}_0^1 = \mathcal{P}_0^0 \oplus \tilde{\mathcal{P}}_0^1$ is unique. This is the smallest space to enforce the continuity of the (constant) tangential component of E across the interelement boundaries.

- It is not necessary but natural to construct the complements using scalar and vector *bubble functions*. In this case, the notation $\tilde{\mathcal{P}}_{-1}^{p+1}$ and $\tilde{\mathcal{P}}_{-1}^p$ is more appropriate. The concept is especially natural if one uses *hierarchical shape functions*. We can always enforce the zero trace condition by augmenting original shape functions with functions of lower order. In other words, we change the complement but *do not alter* the ultimate polynomial space.
- The choice of the complements may be made unique by imposing additional conditions. The original Nedelec's construction, for elements of uniform order p ,

employs symmetric polynomials,

$$\mathcal{R}^p = \{E \in \mathcal{P}^p; \epsilon^p(E) = 0\}$$

where ϵ^p is the Nedelec symmetrization operator,

$$(\epsilon^p(E))_{i_1, \dots, i_{p+1}} = \frac{1}{p+1} \left(\frac{\partial^p E_{i_1}}{\partial x_{i_2} \dots \partial x_{i_p} \partial x_{i_{p+1}}} + \frac{\partial^p E_{i_2}}{\partial x_{i_1} \dots \partial x_{i_p} \partial x_{i_{p+1}}} + \dots + \frac{\partial^p E_{i_{p+1}}}{\partial x_{i_1} \dots \partial x_{i_p} \partial x_{i_{p+1}}} \right)$$

The algebraic complement can then be selected as the subspace of homogeneous symmetric polynomials \mathcal{D}^p [1],

$$\mathcal{R}^p = \mathcal{P}^p \oplus \mathcal{D}^p$$

The space \mathcal{D}^p can be nicely characterized as the image of homogeneous polynomials of order $p-1$ under the Poincaré map; see Hiptmair (1999, 2000),

$$\begin{aligned} E_1(x) &= -x_2 \int_0^1 t \psi(tx) dt \\ E_2(x) &= x_1 \int_0^1 t \psi(tx) dt \end{aligned} \quad (14)$$

The Poincaré map is a right inverse of the curl map, $\nabla \times E = \psi$, for the E defined above. Consistent with our discussion, it can be shown that the tangential trace of a symmetric polynomial of order p is always a polynomial of order less than or equal to $p-1$. The uniqueness of the spaces could also be naturally enforced by requesting *orthogonality* of algebraic complements,

$$\begin{aligned} \mathcal{P}_{p_r}^{p+1} &= \mathcal{P}_{p_r}^p \oplus \tilde{\mathcal{P}}_{-1}^{p+1}, & \mathcal{P}_{-1}^{p+1} &= \mathcal{P}_{-1}^p \oplus \tilde{\mathcal{P}}_{-1}^{p+1} \\ \mathcal{P}_{p_r-1}^p &= \mathcal{P}_{p_r-1}^{p-1} \oplus \nabla(\tilde{\mathcal{P}}_{p_r-1}^{p+1}) \oplus \tilde{\mathcal{P}}_{-1}^p, \\ \mathcal{P}_{-1}^p &= \mathcal{P}_{-1}^{p-1} \oplus \nabla(\tilde{\mathcal{P}}_{-1}^{p+1}) \oplus \tilde{\mathcal{P}}_{-1}^p \end{aligned}$$

The orthogonality in the E space is understood in the $H(\text{curl})$ sense, with the corresponding L^2 orthogonality for the gradients.

The 3D construction goes along the same lines but it becomes more technical. The following decompositions are relevant.

$$\begin{aligned} \mathcal{P}_{p_r, p_f+1}^{p+1} &= \mathcal{P}_{p_r, p_f}^p \oplus \tilde{\mathcal{P}}_{-1, p_f+1}^{p+1} \\ \mathcal{P}_{p_r-1, p_f}^p &= \mathcal{P}_{p_r-1, p_f-1}^{p-1} \oplus \nabla(\tilde{\mathcal{P}}_{p_r-1, p_f+1}^{p+1}) \oplus \tilde{\mathcal{P}}_{-1, p_f}^p \\ \mathcal{P}_{p_f}^p &= \mathcal{P}_{p_f-1}^{p-1} \oplus \nabla(\tilde{\mathcal{P}}_{-1, p_f+1}^{p+1}) \oplus \tilde{\mathcal{P}}_{-1}^p \end{aligned}$$

The ultimate sequence looks as follows:

$$\begin{aligned} \mathcal{P}_{p_r, p_f}^p &\xrightarrow{\nabla} \mathcal{P}_{p_r-1, p_f-1}^{p-1} \oplus \tilde{\mathcal{P}}_{-1, p_f}^p \xrightarrow{\nabla \times} \mathcal{P}_{p_f-1}^{p-1} \\ &\oplus \tilde{\mathcal{P}}_{-1}^p \xrightarrow{\nabla \cdot} \mathcal{P}^{p-1} \end{aligned}$$

Referring to Webb (1999) and Demkowicz (2000) for details, we emphasize only that switching to the tetrahedra of the first type in 3D, requires adding not only extra interior bubbles but face bubbles as well.

Prismatic elements

We shall not discuss here the construction of the exact sequences for the prismatic elements. The prismatic element shape functions are constructed as tensor products of triangular element and 1D element shape functions. We can use both Nedelec's triangles for the construction and, consequently, we can also produce two corresponding exact sequences.

Parametric elements

Given a bijective map $x = x_K(\xi)$ transforming master element \hat{K} onto a physical element K , and master element shape functions $\hat{\phi}(\xi)$, we define the H^1 -conforming shape functions on the physical element in terms of master element coordinates,

$$\phi(x) = \hat{\phi}(\xi) = \hat{\phi}(x_K^{-1}(x)) = (\hat{\phi} \circ x_K^{-1})(x)$$

The definition reflects the fact that the integration of master element matrices is always done in terms of master element coordinates and, therefore, it is simply convenient to define the shape functions in terms of master coordinates ξ . This implies that the parametric element shape functions are compositions of the inverse x_K^{-1} and the master element polynomial shape functions. In general, we do not deal with polynomials anymore. In order to keep the exact sequence property, we have to define the $H(\text{curl})$ -, $H(\text{div})$ -, and L^2 -conforming elements consistently with the way the differential operators transform. For gradients, we have

$$\frac{\partial u}{\partial x_i} = \frac{\partial \hat{u}}{\partial \xi_k} \frac{\partial \xi_k}{\partial x_i}$$

and, therefore,

$$E_i = \hat{E}_k \frac{\partial \xi_k}{\partial x_i}$$

For the curl operator, we have

$$\epsilon_{ijk} \frac{\partial E_k}{\partial x_j} = \epsilon_{ijk} \frac{\partial}{\partial x_j} \left(\hat{E}_l \frac{\partial \xi_l}{\partial x_k} \right)$$

$$= \epsilon_{ijk} \frac{\partial \hat{E}_l}{\partial x_j} \frac{\partial \xi_l}{\partial x_k} + \hat{E}_l \underbrace{\epsilon_{ijk} \frac{\partial^2 \xi_l}{\partial x_j \partial x_k}}_{=0} = \epsilon_{ijk} \frac{\partial \hat{E}_l}{\partial \xi_m} \frac{\partial \xi_m}{\partial x_j} \frac{\partial \xi_l}{\partial x_k}$$

But,

$$\epsilon_{ijk} \frac{\partial \xi_m}{\partial x_j} \frac{\partial \xi_l}{\partial x_k} = J^{-1} \epsilon_{nm} \frac{\partial x_i}{\partial \xi_n}$$

where J^{-1} is the inverse jacobian. Consequently,

$$\epsilon_{ijk} \frac{\partial E_k}{\partial x_j} = J^{-1} \frac{\partial x_i}{\partial \xi_n} \left(\epsilon_{nm} \frac{\partial \hat{E}_l}{\partial \xi_m} \right)$$

This leads to the definition of the $H(\text{div})$ -conforming parametric element,

$$H_i = J^{-1} \frac{\partial x_i}{\partial \xi_n} \hat{H}_n$$

Finally,

$$\frac{\partial H_i}{\partial x_i} = \frac{\partial}{\partial x_i} \left(J^{-1} \frac{\partial x_i}{\partial \xi_k} \hat{H}_k \right) = J^{-1} \frac{\partial x_i}{\partial \xi_k} \frac{\partial \hat{H}_k}{\partial \xi_i} + J^{-1} \frac{\partial x_i}{\partial \xi_k} \frac{\partial \hat{H}_k}{\partial \xi_i} = J^{-1} \frac{\partial \hat{H}_k}{\partial \xi_k}$$

which establishes the transformation rule for the L^2 -conforming elements,

$$f = J^{-1} \hat{f}$$

Defining the parametric element spaces W_p , Q_p , V_p , and Y_p using the transformation rules listed above, we preserve for the parametric element the exact sequence (13).

In the case of the *isoparametric element*, the components of the transformation map x_K come from the space of the H^1 -conforming master element,

$$x_j = \sum_k x_{j,k} \hat{\phi}_k(\xi) = \sum_k x_k \phi_k(x)$$

Here $x_{j,k}$ denote the (vector-valued) geometry degrees of freedom corresponding to element shape functions $\phi_k(x)$. By construction, therefore, the parametric element shape functions can reproduce any linear function $a_j x_j$. As they can also reproduce constants, the isoparametric element space of shape functions contains the space of all linear polynomials in x , i.e. $a_j x_j + b$, in mechanical terms – the space of linearized rigid body motions. The exact sequence property implies that the $H(\text{curl})$ -conforming element can reproduce only constant fields, but the $H(\text{div})$ -conforming element, in general, cannot reproduce even constants. This

indicates in particular that, in context of general parametric (nonaffine) elements, [2] unstructured mesh generators should be used with caution (comp. Arnold, Boffi and Falk, 2000). This critique does not apply to (algebraic) mesh generators based on a consistent representation of the domain as a manifold, with underlying global maps parameterizing portions of the domain. Upon a change of variables, the original problem can then be redefined in the reference domain discretized with affine elements; see, for example, Xue and Demkowicz (2002).

4 PROJECTION-BASED INTERPOLATION. DE RHAM DIAGRAM

In the h version of the FEM, once an element space of shape functions $X(K)$ of dimension N has been specified, the classical notion of a finite element (see Ciarlet, 1978) still requires defining the element degrees of freedom. Element d.o.f. are linear functionals, defined on a larger [3] space $\mathcal{X}(K)$,

$$\psi_j: \mathcal{X}(K) \rightarrow \mathbb{R}, \quad j = 1, \dots, N$$

such that their restrictions on the element space of shape functions are linearly independent. The element shape functions $\phi_i \in X(K)$ are then defined as a dual basis to the d.o.f. functionals,

$$\psi_j(\phi_i) = \delta_{ij}, \quad i, j = 1, \dots, N$$

The purpose of introducing the d.o.f. is twofold:

- to enforce the global conformity by equating appropriate d.o.f. for adjacent elements,
- to define the interpolation operator,

$$\Pi: \mathcal{X}(K) \rightarrow X(K), \quad \Pi u = \sum_{i=1}^N \psi_i(u) \phi_i$$

The interpolation operator is then used not only in proofs of convergence but also in practical computations: mesh generation, approximation of initial and Dirichlet boundary conditions data, and so on.

For the p and hp methods, it is more natural to introduce first shape functions and define interpolation operators directly, without reference to any d.o.f.

Shape functions

With infinite precision, the FE solution depends only on the spaces and not on concrete shape functions. The choice of shape functions, however, affects the conditioning of

stiffness matrices and, in presence of round-off error, the ultimate quality of the FE solution. In context of the p - and hp -versions of the FEM, it is natural to request for the shape functions to be *hierarchical*: increasing order p should result in addition of extra shape functions without modifying the existing ones. Although the choice of shape functions is up to a certain point arbitrary, they have to satisfy basic logic requirements resulting from the conformity considerations. For example, for a tetrahedral element, we must satisfy the following conditions.

- **H^1 -conforming elements.** We have vertex, edge, face, and interior ('bubble') shape functions. A vertex shape function vanishes at the remaining vertices. An edge shape function vanishes along the remaining edges (and, therefore, at all vertices as well). A face shape function vanishes over the remaining faces. The interior bubbles vanish over the whole element boundary. The number of shape functions associated with a particular topological entity is equal to the dimension of the corresponding trace space, for example, we have one shape function per vertex, $p_e - 1$ shape functions per edge of order p_e , $(p_f - 2)(p_f - 1)/2$ shape functions per face of order p_f , and $(p - 3)(p - 2)(p - 1)/6$ interior bubbles.
- **$H(\text{curl})$ -conforming elements.** We have edge, face and interior ('vector bubble') shape functions. Tangential traces of an edge shape function vanish along the remaining edges. Tangential traces of a face bubble vanish over the remaining faces and tangential traces of interior bubbles vanish over the whole boundary of the element. The number of shape functions is again equal to the dimension of the corresponding trace space and it depends upon the element kind.
- **$H(\text{div})$ -conforming elements.** We have only face and interior bubble shape functions. Normal traces of a face shape function must vanish over the remaining faces. Normal traces of interior bubbles over the whole element boundary are zero.

Additional restrictions result from the assumption on functions being hierarchical and, for tetrahedral elements, the rotational invariance assumption. The hierarchical assumption implies that vertex shape functions must be linear. An edge shape function that is of order p_e along the edge should be extended to the rest of the element using a polynomial of order p_e , and so on. The rotational invariance assumption implies that H^1 -conforming shape functions should be defined in terms of barycentric coordinates λ_i , whereas the $H(\text{curl})$ -conforming shape functions should be defined in terms of products $\lambda_i \nabla \lambda_j$. Notice that the invariance of $H(\text{curl})$ -conforming shape functions must be

understood consistently with the definition of the $H(\text{curl})$ -conforming parametric element. If $B: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ denotes the affine map that maps vertex e into vertex $e + 1$ (modulo 4), and ϕ_e denotes a shape function corresponding to edge e , we request that

$$\phi_{e+1}(x) = \phi_e(B^{-1}x) \nabla B(x)$$

The idea of H^1 -conforming shape functions was developed in the pioneering work on the p -method by Szabo and his students; see Szabo and Babuska (1991) and the literature therein; compare also the Chapter 5, this Volume by Düster, Szabo, and Rank in this volume.

Shape functions implied by Nedelec's degrees of freedom are neither unique nor hierarchical. The higher-order $H(\text{curl})$ -conforming shape functions were rediscovered in the engineering community a decade later, starting with Lee, Sun and Cendes (1991). Hierarchical $H(\text{curl})$ -conforming shape functions were first introduced in Webb and Forghani (1993) and further developed in Wang and Webb (1997) and Webb (1999). The last contribution contains a detailed discussion on rotational invariance, the possibility of separating the shape functions into gradients of scalar shape functions and ('rotational') shape functions with nonzero curl, as well as the possibility of enforcing partial L^2 -orthogonality in between the shape functions. In Webb (1999), the author points out clearly the fact that the Nedelec's construction of incomplete ('mixed') order tetrahedra is not unique; see also the discussion above and Demkowicz (2000). For a recent work on optimal selection of hierarchical shape functions; see Ainsworth and Coyle (2001, 2003, 2003). The construction of optimal nonhierarchical $H(\text{curl})$ -conforming shape functions of arbitrary order remains a subject of intensive research as well; see for example Graglia, Wilton and Peterson (1997) and Salazar-Palma *et al.* (1998) and the literature therein.

Projection-based interpolation

The idea of projection-based interpolation stems from three assumptions.

- **Locality.** The element interpolant of a function should be defined entirely in terms of the restriction of the function to the element only.
- **Global continuity.** The union of element interpolants should be globally conforming.
- **Optimality.** The interpolation error should behave asymptotically, both in h and p , in the same way as the actual approximation error.

We make the following regularity assumptions.

$$u \in \mathcal{W} := H^{3/2+r}, \quad r > 0$$

$$E \in \mathcal{Q} := H^{1/2+r}(\text{curl})$$

$$= \{E \in H^{1/2+r}, \nabla \times E \in H^{1/2+r}\}, \quad r > 0$$

$$H \in \mathcal{V} := H^r(\text{div}) = \{H \in H^r; \nabla \cdot H \in H^r\}, \quad r > 0$$

with the corresponding norms denoted by $\|u\|_{1/2+r,T}$, $\|E\|_{1/2+r,T}$, $\|H\|_{r,T}$. Here H^r denotes the fractional Sobolev spaces; see for example Adams (1978).

Let us start with the H^1 -conforming interpolation first. Locality and global continuity imply that the interpolant $u^p = \Pi u$ must match interpolated function u at vertices:

$$u^p(v) = u(v) \quad \text{for each vertex } v$$

With the vertex values fixed, locality and global continuity imply that the restriction of the interpolant to an edge should be calculated using the restriction of function u to that edge only. Optimality, in turn, implies that we should use a projection in some 'edge norm',

$$\|u - u^p\|_e \rightarrow \min, \quad \text{for each edge } e$$

By the same argument, we should then use face projections to determine the face interpolants and, finally, project over the element to complete the definition. The choice of element norm is dictated by the problem being solved: the H^1 -norm for elliptic, the $H(\text{curl})$ -norm for the Maxwell problems, and so on. It follows from the optimality condition then that the face and edge norms are implied by the Trace Theorem; see Lions and Magenes (1972) and Buffa and Ciarlet (2001). The H^1 -conforming interpolant $\Pi u \in X(K)$ of function $u \in H^{1/2+r}(T)$, $r > 0$, is formally defined as follows.

$$\begin{cases} u^p(v) = u(v) & \text{for each vertex } v \\ \|u^p - u\|_{0,e} \rightarrow \min & \text{for each edge } e \\ \|u^p - u\|_{1/2,f} \rightarrow \min & \text{for each face } f \\ \|u^p - u\|_{1,T} \rightarrow \min & \text{for element } K \end{cases}$$

$H(\text{curl})$ -conformity involves only the continuity of the tangential component. Consequently, there is no interpolation at vertices, and the interpolation process starts from edges. Given a function $E \in H^{1/2+r}(\text{curl})$, $r > 0$, we define the projection-based interpolant $\Pi^{\text{curl}} E := E^p$ by requesting the conditions,

$$\begin{cases} \|E^p - E\|_{-1,e} \rightarrow \min, & \text{for each edge } e \\ |(\nabla \times E^p) \cdot n_f - (\nabla \times E) \cdot n_f|_{-1/2,f} \rightarrow \min \\ (E^p - E, \nabla_f \phi)_{-1/2,f} = 0 & \text{for every face bubble } \phi, \\ \text{for each face } f \\ \|\nabla \times E^p - \nabla \times E\|_{0,K} \rightarrow \min \\ (E^p - E, \nabla \phi)_{0,K} = 0, & \text{for every element bubble } \phi \end{cases}$$

Here the norms and inner products correspond to spaces $H^{-1}(\epsilon)$ and $H^{-1/2}(f) = (H_0^{1/2}(f))'$.

For completeness, we also record the projection-based interpolation for $H(\text{div})$ -conforming elements. Given a vector-valued function $\mathbf{H} \in \mathbf{H}^r(\text{div})$, $r > 0$, we define the projection-based interpolant $\Pi^{\text{div}} \mathbf{H} := \mathbf{H}^p \in \mathbf{P}_{p_f}^r$ by requesting the conditions,

$$\begin{cases} \|\mathbf{H}^p \cdot \mathbf{n}_f - \mathbf{H} \cdot \mathbf{n}_f\|_{-1/2, f} \rightarrow \min, & \text{for each face } f \\ \|\nabla \circ \mathbf{H}^p - \nabla \circ \mathbf{H}\|_{0, K} \rightarrow \min \\ \mathbf{H}^p - \mathbf{H}, \nabla \times \phi|_{0, K} = 0 & \text{for every element bubble } \phi \end{cases}$$

Notice that the bubble shape functions are defined consistently with the de Rham diagram. The H^1 -bubbles are used to define the $\mathbf{H}(\text{curl})$ -conforming interpolant and the $\mathbf{H}(\text{div})$ -bubbles are used to define the $\mathbf{H}(\text{div})$ -conforming interpolant.

The interpolation is to be performed on the master element. In order to interpolate on the physical element, we use the element map to transfer the interpolated function to the master element first, interpolate on the master element, and then transfer back the interpolant to the physical element [4].

Under some technical conjectures concerning the existence of polynomial extensions, we have the following result for the tetrahedral element T of the second kind; see Demkowicz and Babuška (2003) and Demkowicz and Buffa (2004) [5].

Theorem 1. For each $\epsilon > 0$, there exist constants $C(\epsilon, r)$ dependent upon ϵ and regularity r but independent of polynomial order p such that the following estimates hold.

$$\begin{aligned} \|\mathbf{u} - \Pi \mathbf{u}\|_{1, T} &\leq C \left(\frac{h}{p}\right)^{1/2+r-\epsilon} \|\mathbf{u}\|_{1/2+r, T} \\ \|\mathbf{E} - \Pi^{\text{curl}} \mathbf{E}\|_{\text{curl}, 0, T} &\leq C \left(\frac{h}{p}\right)^{1/2+r-\epsilon} \|\mathbf{E}\|_{\text{curl}, 1/2+r, T} \\ \|\mathbf{F} - \Pi^{\text{div}} \mathbf{F}\|_{\text{div}, 0, T} &\leq C \left(\frac{h}{p}\right)^{r-\epsilon} \|\mathbf{F}\|_{\text{div}, r, T} \end{aligned}$$

Here h denotes the element size and p is the maximum order of polynomials reproduced by the element space of shape functions. The elements may be of variable order. A similar result is expected to hold for the other elements discussed in the article.

With the projection-based interpolation operators in place, we have the following fundamental result.

Theorem 2. The following de Rham diagram commutes.

$$\begin{array}{ccccccc} \mathcal{W} & \xrightarrow{\nabla} & \mathcal{Q} & \xrightarrow{\nabla \times} & \mathcal{V} & \xrightarrow{\nabla \circ} & L^2 \\ \downarrow \Pi & & \downarrow \Pi^{\text{curl}} & & \downarrow \Pi^{\text{div}} & & \downarrow P \\ W_{hp} & \xrightarrow{\nabla} & Q_{hp} & \xrightarrow{\nabla \times} & V_{hp} & \xrightarrow{\nabla \circ} & Y_{hp} \end{array} \quad (15)$$

Here W_{hp} , Q_{hp} , V_{hp} , and Y_{hp} denote the spaces of element shape functions corresponding to any of the discussed parametric elements. Notice that, contrary to the classical approach, the interpolation procedure does not depend upon the element being considered.

The commuting diagram property is crucial in proving a number of fundamental mathematical results, starting with a proof of the discrete compactness property introduced in Kikuchi (1989). The property mimics a corresponding compactness property on the continuous level. Given a sequence of discrete fields E_{hp} that are uniformly bounded in $\mathbf{H}(\text{curl})$ and discrete divergence-free, that is,

$$\|E_{hp}\|_{\text{curl}, 0, \Omega} \leq C, \quad (\mathbf{E}_{hp}, \nabla \phi_{h, p+1})_{0, \Omega} = 0 \quad \text{for every } \phi_{h, p+1}$$

we can extract a subsequence that converges strongly in L^2 to a limit E . The property has been proved for h -extensions in Boffi (2000), Demkowicz, Monk and Schwab (2000), and Monk and Demkowicz (2000) and recently (under an additional conjecture on polynomials, 2D only) for p and hp -extensions in Boffi, Demkowicz and Costabel (2003). The discrete compactness property implies convergence of Maxwell eigenvalues with optimal rates and, in turn, asymptotic stability and optimal convergence for the time-harmonic Maxwell equations; see Demkowicz and Vardapetyan (1998), Monk and Demkowicz (2000), and Boffi (2001).

The importance of the commuting diagram property in context of the original Nedelec's interpolation was first emphasized in Bossavit (1989). Nedelec's interpolation was intended for h -extensions only and it does not make the diagram commute for variable order elements. It also does not yield an interpolant optimal in p . The projection-based interpolation for variable order elements with the corresponding proof of the commuting property was introduced in Demkowicz *et al.* (2000). For a more general definition of commuting quasi-local interpolation operators, see Schöberl (2001).

5 ADDITIONAL COMMENTS

Exterior boundary value problems

Electromagnetic waves do not penetrate only ideal conductors, and the solution to most practical problems

involves modeling of waves in unbounded domains. The formulation of a boundary value problem then includes also the Silver–Müller condition expressing the fact that waves can propagate only toward infinity (a generalization of Sommerfeld radiation condition for the Helmholtz equations). Finite elements cannot be used to discretize such a problem directly. The most natural approach is then to truncate the infinite domain with a truncating surface, and complement the FE discretization within the truncated domain with a special Absorbing Boundary Condition (ABC). The most popular ABC in context of EM waves is the Berenger's Perfectly Matched Layer (PML) approach; see Berenger (1994). Finite elements can also be naturally coupled with Infinite Elements; see Cecot, Demkowicz and Rachowicz (2003). The most mathematically sound approach is to couple the variational formulation within the truncated domain with a weak form of a boundary integral equation (BIE) set up on the truncating surface. A consistent discretization of the BIE requires the use of RWG (Rao–Wilton–Glisson) elements; see Rao, Wilton and Glisson (1982).

Waveguides form an important class of problems defined in unbounded domains. The solution of a waveguide problem leads to a two-dimensional eigenvalue problem defined in either a bounded or an unbounded domain. Waveguide problems naturally lead also to mixed formulation and the de Rham diagram; see Vardapetyan and Demkowicz (2002) and Vardapetyan, Demkowicz and Neikirk (2003).

Iterative and multigrid solvers

Solutions of significant problems requires the use of iterative and multigrid solvers. The construction and analysis of such solvers for Maxwell equations differs significantly from that for elliptic problems, and it must be based again on Helmholtz decomposition and de Rham diagram; see, for example, Hiptmair (1998), Arnold, Falk and Winther (2000), and Gopalakrishnan and Pasciak (2003).

A posteriori error estimation and adaptivity

A truly effective implementation of an FE code must integrate a posteriori error estimation, adaptivity, and multigrid solvers. For examples of such implementations in context of low order elements; see Beck *et al.* (1999) and Haase, Kuhn and Langer (2001), compare also Salazar-Palma *et al.* (1998).

As I mentioned in the abstract, this presentation is biased very much towards hp elements that enable exponential convergence and have constituted my own research for the past decade. Mathematical foundations for analyzing hp convergence for Maxwell equations start with the fundamental contributions of Costabel and Dauge (2000) on regularity of solutions to Maxwell equations. In his fundamental result, Costabel (1990), demonstrated the failure

of standard penalty-based H^1 -conforming elements to converge in the case of problems with singular solutions that are not in H^1 . In their most recent work, though, Costabel and Dauge (2001) modified the penalty term, using a weighted regularization, and proved and demonstrated the possibility of exponential convergence of the modified method. The first person to analyze the p -extensions using Nedelec's elements was Monk (1994); see also Wang, Monk and Szabo (1996). To my best knowledge, the first hp codes, both in 2D (hybrid meshes consisting of both quadrilaterals and triangles) and in 3D (hexahedra) were put together in Rachowicz and Demkowicz (2000); see also Zdenek and Rachowicz (2002). The codes enabled both h and p adaptivity, with the possibility of anisotropic refinements by means of the constrained approximation for one-irregular meshes. Subsequent implementations in 2D were presented in Ainsworth and Coyle (2001) and in Ledger *et al.* (2003). Experimentally obtained exponential convergence rates were reported in Rachowicz, Demkowicz and Vardapetyan (1999) and in Ainsworth and Coyle (2001). An automatic hp -adaptivity for Maxwell equations based on the minimization of the projection-based interpolation error has recently been presented in Demkowicz (2003).

Transient problems and discontinuous Galerkin FE discretizations

A separate development is taking place in the field of Discontinuous Galerkin (DG) methods for Maxwell equations; see, for example, Perugia, Schötzau and Monk (2002). The DG approach is especially well suited for the solution of transient Maxwell equations; see Hesthaven and Warburton (2001).

6 RELATED CHAPTERS

(See also Chapter 5, Chapter 9, Chapter 13 of this Volume; Chapter 22 of Volume 2)

ACKNOWLEDGMENT

The work has been supported by Air Force under Contract F49620-98-1-0255. The author would like to thank Profs. Peter Monk and Luis Garcia-Castillo for numerous discussions on the subject.

NOTES

- [1] A polynomial of order p is homogeneous if it can be represented as a sum of monomials of order p . Equivalently, $u(\xi x_1, \dots, \xi x_n) = \xi^p u(x_1, \dots, x_n)$.

- [2] Note that general quadrilaterals or hexahedra with straight edges are not affine elements.
- [3] The space should include the solution to the boundary value problem.
- [4] Alternatively, one could perform the projection-based interpolation directly on the physical element. The two interpolants, in general, are different.
- [5] The theorem holds under the assumption of existence of polynomial preserving extension operators for 'energy spaces' H^1 , $H(\text{curl})$, $H(\text{div})$. Existence of such operators has been established for 2D spaces and 3D H^1 space, but it remains so far only conjectured for spaces $H(\text{curl})$, $H(\text{div})$ in 3D.

REFERENCES

- Adams H. *Sobolev Spaces*. Academic Press: New York, 1978.
- Ainsworth M and Coyle J. Hierarchic hp -edge element families for Maxwell's equations on hybrid quadrilateral/triangular meshes. *Comput. Methods Appl. Mech. Eng.* 2001; **190**: 6709–6733.
- Ainsworth M and Coyle J. Conditioning of hierarchic p -version Nedelec elements on meshes of curvilinear quadrilaterals and hexahedra. *SIAM J. Numer. Anal.* 2003; **41**(2):731–750.
- Ainsworth M and Coyle J. Hierarchic Finite Element Bases on Unstructured Tetrahedral Meshes. *Int. J. Numer. Methods Eng.* 2003; **58**(14):2103–2130.
- Arnold D, Falk RS and Weather R. Multigrid in $H(\text{div})$ and $H(\text{curl})$. *Numer. Math.* 2000; **85**(2):197–217.
- Arnold D, Buffi A and Falk RS. Quadrilateral $H(\text{div})$ Finite Elements. I.A.N.-C.N.R. 2002; 1283, 1–23. (<http://www.imati.cnr.it/ian/PUBBLICAZIONI/Publicazioni2002.html>).
- Beck R, Deuffhard P, Hipfmaier R, Hoppe RHW and Wohlmuth B. Adaptive multilevel methods for edge element discretizations of Maxwell's equations. *Surv. Methods Ind* 1999; **8**:271–312.
- Béranger J-P. A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.* 1994; **114**:185–200.
- Boffi D. Discrete compactness and Fortin operator for edge elements. *Numer. Math.* 2000; **87**(2):229–246.
- Boffi D. A note on the discrete compactness property and the de Rham diagram. *Appl. Math. Lett.* 2001; **14**(1):33–38.
- Boffi D, Demkowicz L and Costabel M. Discrete Compactness for p and hp 2D Edge Finite Elements. *Mathematical Models and Methods in Applied Sciences*, Vol. 13, No. 11, 2003; 1673–1687.
- Bossavit A. Un nouveau point de vue sur les éléments finis mixtes. *Matapli (bulletin de la Société de Mathématiques Appliquées et Industrielles)* 1989; **20**:23–35.
- Buffi A and Ciarlet P. On traces for functional spaces related to Maxwell's equations. Part I: An integration by parts formula in lipschitz polyhedra. Part II: Hodge decompositions on the boundary of lipschitz polyhedra and applications. *Math. Methods Appl. Sci.* 2001; **24**(9–30):31–48.
- Ciarlet PG. *The Finite Element Method for Elliptic Problems*. North Holland: New York, 1978.
- Cecot W, Demkowicz L and Rachowicz W. An hp -adaptive finite element method for electromagnetics. Part 3: A three-dimensional infinite element for Maxwell's equations. *Int. J. Numer. Methods Eng.* 2003; **57**:899–921.
- Costabel M and Dauge M. Singularities of electromagnetic fields in polyhedral domains. *Arch. Ration. Mech. Anal.* 2000; **151**: 221–276.
- Costabel M. A coercive bilinear form for Maxwell's equations. *Math. Methods Appl. Sci.* 1990; **12**(4):365–368.
- Costabel M and Dauge M. Weighted Regularization of Maxwell Equations in Polyhedral Domains. Preprint 2001 (<http://www.maths.univ-rennes1.fr/costabel/>).
- Demkowicz L and Vardapetyan L. Modeling of electromagnetic absorption/scattering problems using hp -adaptive finite elements. *Comput. Methods Appl. Mech. Eng.* 1998; **152**(1–2): 103–124.
- Demkowicz L, Monk P, Schwab Ch and Vardapetyan L. Maxwell eigenvalues and discrete compactness in two dimensions. *Math. Comput. Appl.* 2000; **40**(4–5):598–605.
- Demkowicz L, Monk P, Vardapetyan L and Rachowicz R. De Rham diagram for hp finite element spaces. *Math. Comput. Appl.* 2000; **39**(7–8):29–38.
- Demkowicz L. Edge finite elements of variable order for Maxwell's equations. In *Proceedings of the 3rd International Workshop, Warnemuende, 20–23 August; Scientific Computing in Electrical Engineering, (Lecture Notes in Computational Science and Engineering 18)*, Springer-Verlag: Berlin, 2000; 15–34.
- Demkowicz L and Babuška I. "Optimal p Interpolation Error Estimates for Edge Finite Elements of Variable Order in 2D". *SIAM J. Numer. Anal.* 2003; **41**(4):1195–1208.
- Demkowicz L and Buffi A. H^1 , $H(\text{curl})$ and $H(\text{div})$ -Conforming Projection-Based Interpolation in Three Dimensions. *ICES Report 04-24*, The University of Texas at Austin, April 2004.
- Demkowicz L. hp -Adaptive finite elements for time-harmonic Maxwell equations. In *Topics in Computational Wave Propagation, Lecture Notes in Computational Science and Engineering*, Ainsworth M, Davies P, Duncan D, Martin P and Rynne B. (eds). Springer-Verlag: Berlin, 2003; 163–199.
- Gopalakrishnan J and Pasciak JE. Overlapping Schwarz preconditioners for indefinite time harmonic Maxwell equations. *Math. Comput.* 2003; **241**:1–15.(electronic).
- Graglia RD, Wilton DR and Peterson AF. Higher order interpolatory vector bases for computational electromagnetics. *IEEE Trans. Antennas Propagat.* 1997; **45**(3):329–342.
- Haase G, Kuhn M and Langer U. Parallel multigrid 3D Maxwell solvers. *Parallel Comput.* 2001; **27**:761–775.
- Hesthaven JS and Warburton T. Nodal high-order methods on unstructured grids: I. Time-domain solution of Maxwell's equations. *J. Comput. Phys.*, 1 September 2002; **181**(1):186–221.
- Hiptmaier R. Multigrid method for Maxwell's equations. *SIAM J. Numer. Anal.* 1998; **36**(1):204–225.
- Hiptmaier R. Canonical construction of finite elements. *Math. Comput.* 1999; **68**:1325–1346.
- Hiptmaier R. *Higher Order Whitney Forms*. Report 156, Sonderforschungsbereich 382, Universität Tübingen: Tübingen, August 2000.
- Kikuchi F. On a discrete compactness property for the Nedelec finite elements. *J. Fac. Sci. Univ. Tokyo Ser. IA Math.* 1989; **36**(3):479–490.
- Ledger PD, Peraire J, Morgan K, Hassa O and Weatherill NP. Efficient highly accurate hp -adaptive finite element computations of the scattering width output of Maxwell's equations. *Int. J. Numer. Methods Fluids*, 2003; **43**:953–978.
- Lee J-H, Sun DK and Cendes ZJ. Full-wave analysis of dielectric waveguides using tangential vector finite elements. *IEEE Trans. Microwave Theory Tech.* 1991; **39**(8):1262–1271.
- Lions JL and Maganes E. *Non Homogeneous Boundary Value Problems and Applications*, vol. 1. Springer-Verlag: Berlin, 1972.
- Monk P. On the p - and hp -extension of Nedelec's curl-conforming elements. *J. Comput. Appl. Math.* 1994; **53**:117–137.
- Monk P and Demkowicz L. Discrete compactness and the approximation of Maxwell's equations in \mathbb{R}^3 . *Math. Comput.* 2000; **70**(234):507–523.
- Nedelec JC. Mixed finite elements in \mathbb{R}^3 . *Numer. Math.* 1980; **35**:315–341.
- Nedelec JC. A new family of mixed finite elements in \mathbb{R}^3 . *Numer. Math.* 1986; **50**:57–81.
- Perugia I, Schötzau D and Monk P. Stabilized interior penalty methods for the time-harmonic Maxwell equations. *Comput. Methods Appl. Mech. Eng.* 2002; **191**(41–42):4675–4697.
- Rachowicz W, Demkowicz L and Vardapetyan L. hp -Adaptive FE modeling for Maxwell's equations. Evidence of exponential convergence. In *ACES '99, Monterey, 16–20 March, 1999*.
- Rachowicz W and Demkowicz L. An hp -adaptive finite element method for electromagnetics. Part I: Data structure and constrained approximation. *Comput. Methods Appl. Mech. Eng.* 2000; **187**(1–2):307–337.
- Rachowicz W and Demkowicz L. An hp -adaptive finite element method for electromagnetics. Part II: A 3D Implementation. *Int. J. Numer. Methods Eng.* 2002; **53**:147–180.
- Rao SM, Wilton DR and Glisson AW. Electromagnetic scattering by surfaces of arbitrary shape. *IEEE Trans. Antennas Propagat.* 1982; **30**:409–418.
- Salazar-Palma M, Sarkar TK, Garcia-Castillo L-E, Roy T and Djordjević A. *Iterative and Self-Adaptive Finite Elements in Electromagnetic Modeling*. Artech House: London, 1998.
- Schöckel J. *Commuting Quasi-Interpolation Operators for Mixed Finite Elements*. Preprint ISC-01-10-MATH, A&M University: Texas, 2001.
- Senior TBA and Volakis JL. *Approximate Boundary Conditions in Electromagnetics*. IEEE Press: New York, 1995.
- Szabo BA and Babuska I. *Finite Element Analysis*. Wiley: New York, 1991.
- Vardapetyan L and Demkowicz L. Full-wave analysis of dielectric waveguides at a given frequency. *Math. Comput.* 2002; **72**:105–129.
- Vardapetyan L, Demkowicz L and Nelskirk D. hp Vector finite element method for eigenmode analysis of waveguides. *Comput. Methods Appl. Mech. Eng.* 2003; **192**:185–201.
- Wang Y, Monk P and Szabo B. Computing cavity modes using the p version of the finite element method. *IEEE Trans. Mag.* 1996; **32**(3):1934–1940.
- Wang J and Webb JP. Hierarchical vector boundary elements and p -adaptation for 3-D electromagnetic scattering. *IEEE Trans. Antennas Propagat.* 1997; **45**(12):1869–1879.
- Webb JP and Forghani B. Hierarchical scalar and vector tetrahedra. *IEEE Trans. Mag.* 1993; **29**(2):1295–1498.
- Webb JP. Hierarchical vector based functions of arbitrary order for triangular and tetrahedral finite elements. *IEEE Antennas Propagat. Magn* 1999; **47**(8):1244–1253.
- Xue D and Demkowicz L. *Geometrical Modeling Package, Version 2.0*. TICAM Report 02–30, The University of Texas at Austin: USA, August 2002.
- Zdunek A and Rachowicz W. A three-dimensional hp -adaptive finite element approach to radar scattering problems. In *Fifth World Congress on Computational Mechanics Vienna, Austria, 7–12 July, 2002*.

FURTHER READING

- Carstensen C, Funken S, Hackbusch W, Hoppe RHW and Monk P. (eds). *Computational Electromagnetics, Lecture Notes in Computational Science and Engineering 28*. Springer-Verlag: Berlin, 2003.
- Colton D and Kress R. *Inverse Acoustic and Electromagnetic Scattering Theory*. Springer-Verlag: Berlin, 1992.
- Cessenat M. *Mathematical Methods in Electromagnetism*. World Scientific: Singapore, 1996.
- Demkowicz L and Pal M. An infinite element for Maxwell's equations. *Comput. Methods Appl. Mech. Eng.* 1998; **164**: 77–94.
- Jin J. *The Finite Element Method in Electromagnetics*. John Wiley & Sons: New York, 1993.
- Kikuchi F. Mixed and penalty formulations for finite element analysis of an eigenvalue problems in electromagnetism. *Comput. Methods Appl. Mech. Eng.* 1987; **64**:509–521.
- Monk P. *Finite Element Methods for Maxwell's Equations*. Clarendon Press: Oxford, 2003.
- Schwab Ch. p and hp -Finite Element Methods. Clarendon Press: Oxford, 1998.
- Silvester PP and Pelosi G. (eds). *Finite Elements for Wave Electromagnetics*. IEEE Press: Piscataway, 1994.
- Silvester PP and Ferrari RL. *Finite Elements for Electrical Engineers* (3rd edn). Cambridge University Press: New York, 1996.
- Vardapetyan L and Demkowicz L. hp -adaptive finite elements in electromagnetics. *Comput. Methods Appl. Mech. Eng.* 1999; **169**:331–344.
- Volakis JL, Chatterjee A and Kempel LC. *Finite Element Method for Electromagnetics*. IEEE Press: New York, 1998.
- Wang J and Ida N. Curvilinear and higher-order edge elements in electromagnetic field computation. *IEEE Trans. Mag.* 1993; **29**:1491–1494.
- Whitney H. *Geometric Integration Theory*. Princeton University Press: Princeton, New Jersey, USA, 1957.

Contents for Volumes 2 and 3

VOLUME 2: SOLIDS AND STRUCTURES

| | | | |
|--|------------|---|------------|
| Contributors to Volume 2 | ix | 3 Computation | 150 |
| Preface | xi | Notes | 164 |
| | | References | 164 |
| 1 Solids and Structures: Introduction and Survey | 1 | 5 Computational Structural Dynamics | 169 |
| <i>René de Borst</i> | | <i>Gregory M. Hulbert</i> | |
| 1 Introduction | 1 | 1 Introduction | 169 |
| 2 Finite Element Methods for Elasticity with Error-controlled Discretization and Model Adaptivity | 5 | 2 Formulation of Equations of Structural Dynamics | 170 |
| <i>Erwin Stein/Marcus Rüter</i> | | 3 Time Integration Algorithms | 175 |
| 1 Introduction | 5 | 4 Linear Structural Dynamics | 181 |
| 2 Nonlinear and Linear Theory of Elasticity | 7 | 5 Nonlinear Dynamics | 184 |
| 3 Variational Problems and Their Finite Element Discretizations | 16 | 6 Practical Considerations for Time Integration Algorithms | 187 |
| 4 Error Estimation and Adaptivity in Linearized Elasticity | 24 | References | 189 |
| 5 Coupled a Posteriori Model and Discretization Error Estimation | 40 | 6 Computational Contact Mechanics | 195 |
| 6 A Posteriori Error Estimation in Finite Elasticity | 46 | <i>P. Wriggers/G. Zavarise</i> | |
| 7 Concluding Remarks | 54 | 1 Introduction | 195 |
| References | 55 | 2 General Overview | 196 |
| 3 Models and Finite Elements for Thin-walled Structures | 59 | 3 Continuum Description of Contact | 198 |
| <i>M. Bischoff/W. A. Wall/K.-U. Bletzinger/E. Ramm</i> | | 4 Contact Discretizations | 206 |
| 1 Introduction | 59 | 5 Solution Algorithms | 214 |
| 2 Mathematical and Mechanical Foundations | 63 | 6 Conclusions | 221 |
| 3 Plates and Shells | 68 | References | 221 |
| 4 Dimensional Reduction and Structural Models | 70 | 7 Elastoplastic and Viscoplastic Deformations in Solids and Structures | 227 |
| 5 Finite Element Formulation | 107 | <i>F. Armero</i> | |
| 6 Concluding Remarks | 128 | 1 Introduction | 227 |
| Acknowledgments | 128 | 2 The Mechanical Problem | 229 |
| References | 128 | 3 Infinitesimal Models of Elastoplasticity | 232 |
| Further Reading | 133 | 4 Finite Deformation Elastoplasticity | 239 |
| Appendix | 134 | 5 Integration Algorithms | 244 |
| 4 Buckling | 139 | 6 Primal Formulations of the Closest-point Projection Equations | 250 |
| <i>Eduard Riks</i> | | 7 Dual Formulations of the Closest-point Projection Equations | 254 |
| 1 Introduction | 139 | 8 Augmented Lagrangian Formulations | 260 |
| 2 Basic Concepts | 140 | 9 Concluding Remarks | 263 |
| | | Acknowledgment | 264 |
| | | References | 264 |

| | | | |
|---|------------|--|------------|
| 8 Crystal Plasticity and Evolution of Polycrystalline Microstructure | 267 | 4 The Hill-Reuss-Voigt Bounds | 411 |
| <i>Christian Miehe/Jan Schotte</i> | | 5 More Refined Micro-Macro Approximations | 412 |
| 1 Introduction | 267 | 6 Computational Homogenization | 413 |
| 2 The Continuum Slip Theory of Crystal Plasticity | 269 | 7 Microgeometrical Manufacturing Idealizations | 414 |
| 3 Stress Update Algorithms in Single Crystal Plasticity | 274 | 8 Numerical Discretization | 416 |
| 4 Variational Formulation of Elastic-Plastic Crystals | 280 | 9 Overall Testing Process: Numerical Examples | 417 |
| 5 Representative Numerical Examples | 283 | 10 Increasing Sample Size | 420 |
| 6 Conclusion | 287 | 11 Hierarchical/Multiscale Interpretations | 421 |
| References | 287 | 12 Closing Comments and Future Directions | 425 |
| | | Notes | 427 |
| | | References | 427 |
| | | Further Reading | 430 |
| 9 Shakedown and Safety Assessment | 291 | 13 Computational Modeling of Damage and Failures in Composite Laminates | 431 |
| <i>Nestor Zouain</i> | | <i>J. N. Reddy/D. H. Robbins, Jr</i> | |
| 1 Introduction | 291 | 1 Objectives and Scope | 431 |
| 2 Basic Notation | 294 | 2 Introduction: the Multiscale Problem | 432 |
| 3 Classical Static and Kinematic Shakedown Formulations | 296 | 3 Laminate Theories and Models | 433 |
| 4 Extremum Principles for Elastic Shakedown | 301 | 4 Review of Literature on Damage and Failures | 437 |
| 5 Discrete Models for Elastic Shakedown | 304 | 5 Coupling Between the Microscale and the Mesoscale | 440 |
| 6 Preventing Alternating Plasticity | 304 | 6 Modeling of Progressive Damage in Composite Laminates | 451 |
| 7 Preventing Simple Mechanisms of Incremental Collapse | 304 | Acknowledgments | 458 |
| 8 Extensions Concerning Hardening and Temperature Effects | 307 | References | 458 |
| 9 Other Extensions of Shakedown Theory | 310 | Further Reading | 460 |
| 10 An Algorithm for Shakedown Analysis | 312 | | |
| 11 Numerical Procedures for Shakedown Analysis | 316 | 14 Computational Modeling of Forming Processes | 461 |
| 12 Applications | 321 | <i>D. Peric/D. R. J. Owen</i> | |
| 13 Conclusions | 330 | 1 Introduction | 461 |
| References | 331 | 2 Continuum Constitutive Modeling | 462 |
| 10 Damage, Material Instabilities, and Failure | 335 | 3 Implicit Finite Element Solution Strategy | 469 |
| <i>René de Borst</i> | | 4 Contact-Friction Modeling | 471 |
| 1 Introduction | 335 | 5 Element Technology | 476 |
| 2 Damage Mechanics | 336 | 6 Refined Constitutive Models for Inelastic Solids at Finite Strains | 481 |
| 3 Material Instabilities and Mesh Sensitivity | 341 | 7 Thermomechanical Coupling | 484 |
| 4 Cohesive-Zone Models | 349 | 8 Adaptive Strategies for Nonlinear Problems | 485 |
| 5 Enhanced Continuum Descriptions | 355 | 9 Explicit Solution Strategies | 491 |
| 6 Discrete Failure Models | 362 | 10 Thin Sheet Forming Operations | 495 |
| 7 Concluding Remarks | 370 | 11 Bulk Forming Operations | 496 |
| References | 370 | 12 Metal Cutting Operations | 502 |
| 11 Computational Fracture Mechanics | 375 | 13 Concluding Remarks | 506 |
| <i>Anthony R. Ingraffea</i> | | Acknowledgments | 507 |
| 1 Introduction | 375 | Note | 507 |
| 2 A Taxonomy of Approaches for Representation of Cracking Processes | 376 | References | 507 |
| 3 Geometrical Representation Approaches | 378 | Further Reading | 510 |
| 4 Nongeometrical Representation Approaches | 394 | | |
| 5 Summary | 400 | 15 Computational Concrete Mechanics | 513 |
| Acknowledgments | 402 | <i>Roman Lackner/Harbert A. Mang/Christian Pichler</i> | |
| References | 402 | 1 Introduction | 513 |
| 12 Homogenization Methods and Multiscale Modeling | 407 | 2 Basic Ingredients of Multiscale Modeling | 514 |
| <i>Tarek I. Zohdi</i> | | 3 Autogenous Shrinkage of Shotcrete in Hybrid Analysis of Tunnel Linings | 516 |
| 1 Introduction | 407 | 4 Tension Stiffening in Cooling Tower Analysis | 529 |
| 2 Fundamental Micro-Macro Concepts | 409 | 5 Concluding Remarks | 538 |
| 3 Testing Procedures | 409 | References | 539 |

| | | | |
|--|------------|---|------------|
| 16 Computational Geomechanics Including Consolidation | 543 | 20 Stochastic Finite Element Methods | 657 |
| <i>John P. Carter/John C. Small</i> | | <i>Miguel A. Gutiérrez/Steen Krenk</i> | |
| 1 Introduction | 543 | 1 Introduction | 657 |
| 2 Characteristics of Geotechnical Problems | 544 | 2 Random Variables | 658 |
| 3 Deterministic Geotechnical Analysis | 544 | 3 Stochastically Determinate Problems | 661 |
| 4 Methods of Numerical Analysis | 545 | 4 Representation of Random Fields | 663 |
| 5 Limit Analysis Using Finite Elements | 549 | 5 Basic Variable Sensitivity Computation | 667 |
| 6 Constitutive Models for Geomaterials | 551 | 6 Perturbation Technique | 668 |
| 7 Soil-Structure Interaction | 558 | 7 Spectral Formulation | 671 |
| 8 Consolidation | 561 | 8 Reliability Methods | 676 |
| 9 Stochastic Techniques | 567 | 9 Status and Outlook | 680 |
| 10 Concluding Remarks | 569 | References | 680 |
| Acknowledgments | 569 | | |
| References | 569 | 21 Fluid-structure Interaction Problems | 683 |
| | | <i>Roger Ouyon</i> | |
| 17 Multifield Problems | 575 | 1 Introduction | 683 |
| <i>B. A. Schrefler</i> | | 2 Structural-acoustic Problem | 684 |
| 1 Introduction | 575 | 3 Structural-acoustic Equations | 684 |
| 2 Partitioned Solution Procedures | 577 | 4 Incompressible Hydroelastic-clothing Problem | 689 |
| 3 Soil Dynamics | 589 | 5 Hydroelastic-clothing Equations | 689 |
| 4 Monolithic Solution Procedure | 592 | 6 Conclusion | 692 |
| 5 Conclusions | 599 | References | 692 |
| Acknowledgments | 599 | Further Reading | 693 |
| References | 599 | | |
| Further Reading | 603 | 22 Acoustics | 695 |
| | | <i>Lanny L. Thompson/Peter M. Pinsky</i> | |
| 18 Computational Biomechanics of Soft Biological Tissue | 605 | 1 Introduction | 695 |
| <i>Gerhard A. Holzapfel</i> | | 2 Acoustic Field Equations | 696 |
| 1 Introduction | 605 | 3 Time-harmonic Waves and the Helmholtz Equation | 697 |
| 2 Mechanics of the Arterial Wall | 606 | 4 Discretization Methods for the Helmholtz Equation | 698 |
| 3 Mechanics of the Heart Wall | 618 | 5 The Exterior Problem in Unbounded Domains | 702 |
| 4 Mechanics of the Ligament | 625 | 6 The D/N Nonreflecting Boundary Condition | 703 |
| Acknowledgments | 629 | 7 The Modified D/N Nonreflecting Boundary Condition | 705 |
| References | 630 | 8 Infinite Elements | 708 |
| | | 9 Perfectly Matched Layer (PML) | 710 |
| 19 Identification of Material Parameters for Constitutive Equations | 637 | 10 Accelerated Multifrequency Solution Methods | 710 |
| <i>R. Mahnen</i> | | 11 Parallel Iterative Solution Methods | 711 |
| 1 Introduction | 637 | 12 Domain Decomposition Methods | 712 |
| 2 General Framework for Development of Constitutive Models | 638 | 13 Direct Time-domain Methods for Acoustic Waves | 713 |
| 3 Parameter Identification | 640 | 14 Conclusions | 713 |
| 4 Identification Methods | 641 | 15 Related Chapters | 714 |
| 5 Optimization Methods | 643 | Acknowledgments | 714 |
| 6 Instabilities in Least-squares Problems | 645 | References | 714 |
| 7 Stochastic Methods | 646 | | |
| 8 Parameter Identification for Uniform Small Strain Problems | 647 | 23 Boundary Integral Equation Methods for Elastic and Plastic Problems | 719 |
| 9 Parameter Identification for Nonuniform Large Strain Problems | 649 | <i>Marc Bonnet</i> | |
| 10 Concluding Remarks | 652 | 1 Introduction | 719 |
| Acknowledgment | 652 | 2 Basic Integral Identities | 720 |
| References | 654 | 3 The Boundary Element Method in Elasticity: Collocation | 723 |
| Further Reading | 655 | 4 The Boundary Element Method in Elasticity: Symmetric Galerkin | 727 |
| | | 5 Fast Solution Techniques | 729 |
| | | 6 The Boundary Element Method for Fracture Mechanics | 732 |
| | | 7 Boundary-domain Integral Equations for Elastic-Plastic Problems | 735 |
| | | 8 Shape Sensitivity Analysis | 740 |

| | | | |
|--|------------|--|------------|
| 9 FEM-BEM Coupling | 744 | 2 Dual Reciprocity Method for Elastodynamics and Piezoelectricity | 758 |
| 10 Related Chapters | 745 | 3 Nonsingular Hybrid Boundary Element Formulation for Elastodynamics | 762 |
| References | 745 | References | 768 |
| 24 Boundary Element Methods for the Dynamic Analysis of Elastic, Viscoelastic, and Piezoelectric Solids | 751 | Contents for Volumes 1 and 3 | 771 |
| <i>L. Gaul/M. Kögler/F. Moser/M. Schanz</i> | | Subject Index | 777 |
| 1 Viscoelastic Direct Boundary Element Formulation in Time Domain | 751 | | |

VOLUME 3: FLUIDS

Contributors to Volume 3

Preface

| | | | |
|--|-----------|---|------------|
| 1 Fluids: Introduction and Survey | 1 | 5 Vortex Methods | 129 |
| <i>Thomas J. R. Hughes</i> | | <i>G. S. Winckelmans</i> | |
| 2 Multiscale and Stabilized Methods | 5 | 1 Introduction | 129 |
| <i>Thomas J. R. Hughes/Guglielmo Scovazzi/Leopoldo P. Franca</i> | | 2 Vortex Methods for 2D Unbounded Inviscid Flows | 131 |
| 1 Introduction | 5 | 3 Vortex Methods for 3D Unbounded Inviscid Flows | 134 |
| 2 Dirichlet-to-Neumann Formulation | 8 | 4 Viscous Flows | 137 |
| 3 Variational Multiscale Method | 11 | 5 Improvement of the Truncation Error | 139 |
| 4 Space-Time Formulations | 27 | 6 Particle Redistribution | 140 |
| 5 Stabilized Methods for Advection-Diffusive Equations | 32 | 7 Efficient Velocity Evaluation: Fast Multipole Methods | 141 |
| 6 Turbulence | 40 | 8 Efficient Velocity Evaluation: Vortex-in-cell Methods | 144 |
| Acknowledgments | 55 | 9 Flows with Solid Boundaries | 145 |
| References | 55 | 10 Other Applications | 149 |
| 3 Spectral Element and hp Methods | 61 | 11 Related Chapters | 152 |
| <i>Robert M. Kirby/George Em Karniadakis</i> | | Acknowledgments | 152 |
| 1 Introduction | 61 | References | 152 |
| 2 Polynomial Expansions on Unstructured Grids | 63 | Further Reading | 153 |
| 3 Incompressible Flows | 67 | 6 Incompressible Viscous Flows | 155 |
| 4 Compressible Flows | 74 | <i>Rolf Rannacher</i> | |
| 5 Plasma Flows | 81 | 1 Introduction | 155 |
| 6 Discussion | 85 | 2 Mathematical Models | 156 |
| 7 Related Chapters | 88 | 3 Discretization of Space | 160 |
| Acknowledgments | 88 | 4 Discretization of Time | 166 |
| References | 88 | 5 Error Control and Mesh Adaptation | 170 |
| 4 Discontinuous Galerkin Methods for Computational Fluid Dynamics | 91 | 6 Solution of the Algebraic Systems | 175 |
| <i>B. Cockburn</i> | | Acknowledgment | 179 |
| 1 Introduction | 91 | References | 179 |
| 2 Linear Hyperbolic Problems | 92 | 7 Computability and Adaptivity in CFD | 183 |
| 3 Nonlinear Hyperbolic Problems | 96 | <i>J. Hoffman/C. Johnson</i> | |
| 4 DG Methods for Second-order Elliptic Problems | 115 | 1 Introduction | 183 |
| 5 DG Methods for Convection-dominated Problems | 120 | 2 Outline | 186 |
| | | 3 References | 186 |
| | | 4 The Incompressible Navier-Stokes Equations | 186 |
| | | 5 Discretization: General Galerkin G^2 | 187 |

| | | | |
|---|------------|--|------------|
| 6 Adaptive Computation of the Drag of a Bluff Body | 189 | 5 Discretization Scheme for Flows in Complex Multidimensional Domains | 359 |
| 7 Drag of a Square Cylinder | 191 | 6 Time-stepping Schemes | 365 |
| 8 The Drag of a Surface-mounted Cube | 192 | 7 Aerodynamic Shape Optimization | 379 |
| 9 The Drag Versus the Total Dissipation | 195 | 8 Related Chapters | 400 |
| 10 Reliability and Efficiency of the Adaptive Method | 197 | Acknowledgment | 400 |
| 11 Averaged Navier-Stokes Equations and Reynolds Stresses | 198 | References | 400 |
| 12 The Subgrid Model from Stabilization | 198 | 12 Industrial Aerodynamics | 407 |
| 13 Computability in Transition to Turbulence | 199 | <i>Frédéric L. Chalot</i> | |
| 14 Applications to Stationary Benchmark Problems in 3D | 201 | 1 Introduction | 407 |
| 15 Related Chapters | 205 | 2 The Historical Development of Computational Flow Mechanics | 408 |
| Acknowledgments | 205 | 3 Euler codes | 413 |
| References | 205 | 4 Navier-Stokes Code | 416 |
| 8 Dynamic Multilevel Methods and Turbulence | 207 | 5 Reynolds-averaged Turbulence Modeling | 418 |
| <i>T. Dubois/F. Jauberteau/R. M. Temam</i> | | 6 Large Eddy Simulation | 423 |
| 1 Introduction | 207 | 7 Mesh Generation | 426 |
| 2 Turbulence: Multilevel Methods for the Navier-Stokes Equations | 209 | 8 Validation | 428 |
| 3 Multilevel Methods for the Shallow Water Equations | 240 | 9 Code Interface | 428 |
| 4 Renormalization of Small Eddies | 259 | 10 Military Applications | 428 |
| 5 Conclusion: Summary and Perspectives | 263 | 11 Civil Applications | 431 |
| Acknowledgments | 264 | 12 Space Applications | 433 |
| References | 264 | 13 Fundamental Studies and Research Work | 436 |
| 9 Turbulence Direct Numerical Simulation and Large-eddy Simulation | 269 | 14 Shape Optimization | 437 |
| <i>Pierre Sagaut</i> | | 15 Aerodynamics | 444 |
| 1 Introduction | 269 | 16 Multidisciplinary Applications | 448 |
| 2 Mathematical Models and Governing Equations | 271 | 17 Conclusion | 453 |
| 3 Basic Numerical Issues for DNS and LES | 279 | 18 Related Chapters | 454 |
| 4 Subgrid-scale Modeling for the Incompressible Case | 284 | Acknowledgments | 454 |
| 5 Extension of Subgrid Models for the Compressible Case | 290 | References | 454 |
| 6 Boundary Conditions for LES | 291 | Further Reading | 457 |
| 7 Applications of DNS and LES | 293 | 13 CFD-based Nonlinear Computational Aeroelasticity | 459 |
| 8 Related Chapters | 296 | <i>Charbel Farhat</i> | |
| Acknowledgment | 296 | 1 Introduction | 459 |
| References | 296 | 2 The Three-field Formulation of Nonlinear Aeroelastic Problems | 461 |
| 10 Turbulence Closure Models for Computational Fluid Dynamics | 301 | 3 Recent Computational Advances | 464 |
| <i>Paul A. Durbin</i> | | 4 The Aero Simulation Platform | 473 |
| 1 Introduction: The Role of Statistical Closure | 301 | 5 Sample Recent Validation Results | 474 |
| 2 Reynolds-averaged Navier-Stokes Equations | 301 | 6 Conclusions | 476 |
| 3 Models With Scalar Variables | 303 | Acknowledgments | 477 |
| 4 Second Moment Transport | 311 | References | 477 |
| 5 Reynolds-averaged Computation | 318 | 14 The Use of Mixed Finite Element Methods for Viscoelastic Fluid Flow Analysis | 481 |
| Note | 322 | <i>Frank P. T. Baaijens/Martien A. Hulsen/Patrick D. Anderson</i> | |
| References | 322 | 1 Introduction | 481 |
| 11 Aerodynamics | 325 | 2 Mathematical Formulation | 481 |
| <i>Antony Jameson</i> | | 3 Steady Flow: Variational Formulations | 482 |
| 1 Focus and Historical Background | 325 | 4 Time Dependent Flows | 488 |
| 2 Mathematical Models of Fluid Flow | 330 | 5 Integral and Stochastic Constitutive Models | 490 |
| 3 Potential Flow Methods | 334 | 6 Governing Equations | 491 |
| 4 Shock-capturing Algorithms for the Euler and Navier-Stokes Equations | 348 | 7 The Deformation Fields Method | 492 |
| | | 8 Brownian Configuration Fields | 493 |
| | | 9 Numerical Methods | 494 |
| | | 10 Results | 494 |
| | | 11 Conclusions and Discussion | 495 |

| | | | |
|---|------------|--|------------|
| 12 Related Chapters References | 496 496 | 13 Enhanced-Discretization Interface-Capturing Technique (EDICT) | 560 560 |
| 15 Combustion | 499 | 14 Extensions and Offshoots of EDICT | 560 |
| <i>T. J. Poinso/D. P. Veynante</i> | | 15 Mixed Interface-Tracking/Interface-Capturing Technique (MITICT) | 561 |
| 1 Introduction | 499 | 16 Edge-Tracked Interface Locator Technique (ETILT) | 562 |
| 2 Combustion Regimes | 500 | 17 Line-Tracked Interface Update Technique (LTIUT) | 564 |
| 3 Governing Equations | 504 | 18 Iterative Solution Methods | 565 |
| 4 Combustion Terminology and Basics | 507 | 19 Enhanced Solution Techniques | 566 |
| 5 Homogeneous Reactors and Laminar Flames | 513 | 20 Mixed Element-Matrix-Based/Element-Vector-Based Computation Technique (MMVCT) | 567 |
| 6 Turbulent Flames | 514 | 21 Enhanced-Discretization Successive Update Method (EDSUM) | 568 |
| 7 Conclusions | 523 | 22 Examples of Flow Simulations | 570 |
| 8 Related Chapters | 523 | 23 Concluding Remarks | 574 |
| References | 523 | 24 Related Chapters | 574 |
| | | Acknowledgment | 574 |
| | | References | 574 |
| 16 Blood Flow | 527 | 18 Ship Hydrodynamics | 579 |
| <i>Charles A. Taylor</i> | | <i>Eugenio Ohate/Julio Garcia/Sergio R. Idelsohn</i> | |
| 1 Introduction | 527 | 1 Introduction | 579 |
| 2 Overview of the Cardiovascular System | 529 | 2 The Navier-Stokes Equations for Incompressible Flows. ALE Formulation | 581 |
| 3 Lumped Parameter Models | 530 | 3 About the Finite Element Solution of the Navier-Stokes Equations | 583 |
| 4 One-dimensional Wave Propagation Models | 531 | 4 Basic Concepts of the Finite Calculus (FIC) Method | 584 |
| 5 Three-dimensional Equations of Blood Flow | 533 | 5 FIC Equations for Viscous Incompressible Flow. ALE Formulation | 585 |
| 6 Future Work | 540 | 6 Finite Element Discretization | 587 |
| Acknowledgments | 540 | 7 Fluid-ship Interaction | 589 |
| References | 540 | 8 A Simple Algorithm for Updating the Mesh Nodes | 590 |
| 17 Finite Element Methods for Fluid Dynamics with Moving Boundaries and Interfaces | 545 | 9 Modeling of the Transom Stern Flow | 591 |
| <i>Tayfun E. Tezduyar</i> | | 10 Lagrangian Flow Formulation | 591 |
| 1 Introduction | 545 | 11 Modeling the Structure as a Viscous Fluid | 592 |
| 2 Governing Equations | 547 | 12 Computation of the Characteristic Lengths | 592 |
| 3 Stabilized Formulations | 548 | 13 Turbulence Modeling | 593 |
| 4 DSD/SST Finite Element Formulation | 549 | 14 Examples | 594 |
| 5 Calculation of the Stabilization Parameters for Incompressible Flows | 549 | 15 Concluding Remarks | 607 |
| 6 Discontinuity-Capturing Directional Dissipation (DCDD) | 550 | 16 Related Chapters | 607 |
| 7 Calculation of the Stabilization Parameters for Compressible Flows and Shock-Capturing | 551 | Acknowledgments | 607 |
| 8 Mesh Update Methods | 553 | References | 607 |
| 9 Shear-Slip Mesh Update Method (SSMUM) | 555 | | |
| 10 DSD/SST Formulation for Fluid-Object Interactions in Spatially Periodic Flows | 555 | | |
| 11 Space-Time Contact Technique (STCT) | 557 | | |
| 12 Fluid-Object Interactions Subcomputation Technique (FOIST) | 558 | | |

Contents for Volumes 1 and 2

Subject Index

611

619

Subject Index

a posteriori error estimates
BEM/FEM 1:377, 1:382–389, 1:401–402, 1:407–408
buff body drag 3:184, 3:191
coarse scales 3:17–v3:18, 3:19–v3:20
conservation laws 1:448–v1:449
elliptic boundary values 1:95–v1:97
finite elasticity 2:5–7, 2:46–54
finite elements 1:85–98, 2:5–7, 2:46–54
finite hyperelasticity 2:46–54
goal-orientated 2:26–27, 2:33–37, 2:48–50
H-version symmetric BEM/FEM 1:382–389
heat equation 1:690–691
incompressible viscous flows 3:171–173
linearized elasticity 2:27–37, 2:40–46
parabolic differential equations 1:677–678, 1:690–691, 1:695–696
Signorini-type interfaces 1:401–402
a priori error estimates
conservation laws 1:449
convergence characteristics 1:127
finite element methods 1:73, 1:82–85, 1:104–107, 1:127
finite volume elements 1:449, 1:456
H¹-norm 1:733, 2:25–27
heat equation 1:690–691
indefinite finite elements 1:104–105
L₂-norm 2:27–28, 3:261–262
linearized elasticity 2:25–27
nonconforming finite elements 1:105–107
nonsymmetric finite elements 1:104–105
parabolic differential equations 1:690–691, 1:692–695
two-dimensional 1:83–84
a priori mesh design 1:98–99
ABC see Absorbing Boundary Condition
abdominal aorta 3:535, 3:536–537
abrasive wear 2:206
absolute errors 3:96, 3:97
Absorbing Boundary Condition (ABC) 1:735, 2:705
abstract Dirichlet boundary conditions 3:12–18
abstract estimates 1:75–76
ACA see adaptive cross approximations
accelerated multifrequency methods 2:710–711
acceleration
acoustic field equations 2:697
arbitrary Lagrangian-Eulerian methods 1:419
convergence 1:654–655
structural dynamics 2:173
accuracy
critical points 1:453
time integration 2:178, 2:187–188
turbulence closure 3:320–322
ACL see anterior cruciate ligament
acoustics 2:695–714
accelerated multifrequency methods 2:710–711
aeroacoustics 3:444–448
aerodynamic flow mechanics 3:444–448
direct time-domain methods 2:713
Dirichlet-to-Neumann boundary conditions 2:696, 2:703–708, 2:711
domain decomposition 2:712–713
eddy viscosity 3:50–52
environmental 3:5–7
field equations 2:696–697
flame interactions 3:503–504
Helmholtz equation 2:697–702
high-speed trains 3:5–7
infinite elements 2:708–710, 2:711–712
jet temperature 3:445–446
parallel iteration 2:711–712
perfectly matched layers 2:696, 2:710
structural 2:683, 2:684–689
tensors 2:343–344
time-harmonic waves 2:697–698
unbounded domains 2:702–703
acquired cardiovascular disease 3:528
active...
arterial wall mechanics 2:608
cardiac muscle mechanics 2:621–622
set strategy 2:202, 2:203, 2:216
adaptive...
boundary element methods 2:386–390
computation 1:6
averaged Navier-Stokes equations 3:193
convection-diffusion reactions 1:696–699, 1:700
drag of buff bodies 3:189–191
efficiency 3:197–198
error estimation 1:87, 1:97, 1:690–691
heat equations 1:676, 1:689–692, 1:696–699, 1:714–719
literature 1:676–678
nonsliff initial value problems 1:680–683
parabolic differential equations 1:675–702
reaction-diffusion equations 1:696–699, 1:700
reliability 3:197–198
software 1:702
square cylinder drag 3:191–192, 3:193–194, 3:197–198
stationary benchmark problems 3:201–205
stiff initial value problems 1:680
strong stability factors 1:675–679, 1:683, 1:686–689, 1:691–693
surface-mounted cube drag 3:192–195, 3:196, 3:197–198
turbulence 3:199–201
cross approximations (ACA) 2:731
direct numerical simulations 3:184–185, 3:191–197
error control 1:677–680

adaptive... (continued)
 finite element methods
 convergence 1:58, 1:103
 forming processes modeling 2:485–491
 fracture mechanics 2:386–390
 mesh generation/adaptivity 1:510–516
 hierarchical modeling 2:425
 large eddy simulations 3:184–185, 3:191–197
 meshes
 coarsening 1:104
 error estimation 1:87, 1:97
 linearized elasticity 2:37–39
 local refinements 1:73, 1:98, 1:99–100
 spatial representation 1:529–530
 partitions 1:171–172
 space-time Galerkin FEM 1:676, 1:689–690
 support stencil reconstructions 1:462–463
 time steps 1:696–697
 wavelet techniques 1:3, 1:157–195
 boundary integral equations 1:175–181
 complexity analysis 1:187–195
 eddy viscosity 3:49–50
 evolution equations 1:169–175
 numerical simulations 1:157–195
 operators 1:175–181
 residual approximations 1:187–195
 adaptivity 1:1
 buckling computation 2:152–153
 computational fluid dynamics 3:2, 3:183–205
 convection–diffusion equations 1:40–43
 convergence 1:103
 discontinuous Galerkin methods 3:101–102, 3:114–115, 3:116
 error estimates 1:511–516
 linearized elasticity 2:24–40
 Maxwell equations 1:735
 operators 1:160, 1:165, 1:175–181
 quality meshing 1:502–510
 regular meshes 1:502–504
 RKDG 3:114–115, 3:116
 shock capturing 3:101–102
 symmetric BEM/FEM 1:382–389
 addition hierarchical matrices 1:611
 Additive Schwarz Method (ASM) 1:390–391, 1:623–625, 1:627–629, 1:632
 additive variants 1:589–590
 adhesion 2:205
 adhesive wear 2:206
 adjacent equilibrium states 2:142
 adjoint methods 2:742–743, 3:381–386, 3:390
 admissibility 1:601–602, 1:606, 1:608–609, 2:582
 advancing-fronts 1:497–498, 1:500–501, 1:508
 advection
 arbitrary Lagrangian–Eulerian methods 3:76–77
 diffusion equations 3:22–26, 3:32–40, 3:548
 diffusion operators 3:118
 equations 1:148–150
 hp-finite methods 3:62
 shallow water equations 3:250–251
 AERO simulation platforms 3:473–477
 aeroacoustics 3:444–448
 aerodynamics 3:2–3, 3:325–400, 3:407–454
 aerocoustics 3:444–448
 aeroelasticity 3:459–477
 afterbody design 3:451–452
 air conditioning 3:452–453

antenna integration 3:451
 classical 3:325–326
 complex geometry 3:346–348
 computational flow mechanics
 aerocoustics 3:444–448
 code interfaces 3:428
 Euler codes 3:413–416
 flutter 3:448–449, 3:450
 fundamental studies 3:456–457
 historical overviews 3:408–413
 large eddy simulations 3:423–426
 mesh generation 3:426–428
 Navier–Stokes code 3:416–418
 numerical codes 3:413–418
 parafolds 3:449–451
 research 3:436–437
 Reynolds averaged turbulence 3:418–423
 shape optimization 3:437–443
 validation 3:428
 electromagnetism 3:451
 finite difference solvers 3:409–411
 fluid flow discretization 3:359–365
 historical overviews 3:325–329
 industrial 3:407–454
 infrared signatures 3:451–452
 mathematical models 3:330–334, 3:335
 nonlinear aeroelasticity 3:459–477
 planforms 3:392–396
 potential flow methods 3:334–348
 shape optimization 3:379–400
 case studies 3:390–399
 Euler equations 3:383–386
 Navier–Stokes equations 3:383–386
 shock capturing 3:329, 3:348–359, 3:363–364, 3:460
 structural optimization 3:392–396
 subsonic linearized potential flow 3:334–337
 time-stepping schemes 3:365–379
 vortex methods 3:145–149
 aeroelasticity 3:459–477
 Aerospace Research and Development (AGARD) 3:377–379, 3:474, 3:449–450
 aerothermal design 3:433–434
 AETHER 3:416–418
 affine
 interpolation-equivalent elements 1:80–81
 invariance 1:657–658
 mapping 1:58, 1:62–63
 afterbody design 3:451–452
 AGARD see Aerospace Research and Development
 aggregate Interlock 2:536–537
 aggregate scale 2:520–521, 2:524
 AIAA see American Institute of Aerodynamics and Astronauts
 air conditioning 3:452–453
 air flow 2:584–586
 Air Force Office, Scientific Research 3:473–477
 airframe efficiency 3:432–433
 aircraft see aerodynamics
 airfoils
 aeroelasticity 3:460
 complex geometry 3:346
 compressible flows 3:77–81
 NACA 6 series 3:326–327
 NACA 0012 3:78–81, 3:101–103, 3:122–123
 NACA 0015 3:77–78
 airframe noise 3:444

Akzo–Nobel system 1:680, 1:682, 1:685, 1:687
 ALE see arbitrary Lagrangian–Eulerian
 algebraic...
 complements 1:730–731
 expansions 1:145–146
 operators 3:30–32
 polynomial expansions 1:143–145
 solution algorithms 3:175–179
 space discretization 3:166
 algorithms
 Bi-CG 1:559–560
 Bi-CGSTAB 1:559–560, 2:707
 bisection 1:101–102
 Bowyer–Watson 1:501
 buckling stability 2:156
 closest-point-projection equations 2:262–263
 consistency tangents 2:248
 Constraint Energy Momentum 2:186–187
 contact mechanics 2:197, 2:198, 2:214–221
 continuum mechanics 1:413–414
 cutting-plane algorithms 2:246
 kinematic shake-down 2:318–319
 Newton–Raphson 2:18, 2:471
 QMR 1:559–560, 2:707
 radial return 2:737–738
 Red–Green–Blue refinement closure 1:102–103
 shake-down 2:312–321
 visualization 1:531–541
 aliasing errors 3:282, 1:149
 Almansi strain tensors 2:9–10, 1:136
 alternating direct implicit (ADI) methods 1:27–28
 alternating plasticity 2:304, 2:321–322
 Alternating Schwarz Method 1:618, 1:620
 alumina plates 2:379
 aluminum–boron 2:417–421
 aluminum–magnesium 2:648–649
 American Cup Bravo España 3:600, 3:601
 American Cup Rioja de España 3:596–601
 American Institute of Aerodynamics and Astronauts (AIAA) 3:327
 Ampère's law 1:723–724
 anchor bolts 2:360–361
 aneurysms models 2:617–618
 angioplasty 2:617, 2:618
 angle of attack 3:460, 3:429–431, 3:436, 3:443
 anisotropic...
 bound meshes 1:513–514
 damage models 2:338–339
 elastodynamics 2:759–760
 elements 1:57, 1:63–66
 elliptic partial differential equations 1:14
 error estimates 1:81–82
 finite elements 2:20–21
 h-finite elements 1:57, 1:63–66
 hardening plasticity 2:553
 hp-adaptivity refinement 2:38–39
 meshes 1:513–514, 3:114–115, 3:116, 3:163–164
 nodal interpolation error estimates 1:63–66
 Sobolev norms 1:710, 1:711
 turbulence closure 3:311
 annihilating radial terms 2:705
 antenna integration 3:451
 anterior cruciate ligament (ACL) 2:626
 aero-aero bypass models 3:539–540
 approximations
 inertial manifolds 3:211–212
 levels 3:269–270

meshfree methods 1:280–291
 operators 1:88
 plates 1:207
 Riemann solvers 3:97, 3:98–100
 shells 1:218
 arbitrary crack growth 2:382–394
 arbitrary Lagrangian–Eulerian (ALE) methods 1:4, 1:413–433
 advection 3:76–77
 aeroelasticity 3:464–466
 fluid dynamics 1:421, 1:422–426
 kinematics 1:416–417
 mesh-updating 1:420–422
 nonlinear solid mechanics 1:426, 1:428–433
 ship hydrodynamics 3:581–583, 3:585–587, 3:592–593, 3:594–607
 solid mechanics 1:426, 1:428–433
 arbitrary shape approaches 2:382–394
 arch structures 2:558–561
 Argyris elements 1:77–78, 1:85
 Arnold–Winther elements 1:269
 Arnoldi method 1:556–557, 1:571–572
 Arhenius laws 3:517
 arterial...
 aneurysms 2:617–618
 bifurcations 2:617, 2:618, 3:537–538
 blood flow 3:530
 clamping 2:617
 wall mechanics 2:606–618
 artificial boundaries 2:703–704, 2:708, 3:6–7, 3:159–160
 artificial diffusion 3:449–351, 3:583
 artificially cemented sand 2:555–558, 2:559
 ASM see Additive Schwarz Method
 associated plasticity 2:237–239
 assumptions on the norm 1:247–248
 Astley–Lais conjugated test functions 2:709
 asymmetrical mechanical part forming 1:518, 1:519
 asymptotic...
 consistency 1:207, 1:218–219
 decomposition preconditioners 1:619
 error estimates 1:370–371
 exactness 1:86
 expansions 1:201–207, 1:211–218
 atomistic methods 2:383, 2:391–394
 attachment points 1:536
 attraction basins 1:653–654
 attributes, geometric modeling 1:490–492
 Aubin–Nitsche lemma 1:364–365
 augmented...
 closest-point-projections 2:262–263
 consistency functions 2:261–262
 dual functions 2:261–262
 functionals 1:267
 Lagrangian formulations
 contact mechanics 2:202, 2:217–220
 discontinuous deformations 1:329
 elastoplastic deformations 2:260–263
 multibody contact forces 1:322, 1:329
 viscoplastic deformations 2:260–263
 austenitic steel 2:649, 2:650
 auto-catalytic reactions 1:698–699, 1:700
 autogenous shrinkage 2:514, 2:516–529
 autoignition 3:511
 automatic...
 aerodynamic shape optimization 3:437–443
 coarsening 1:104
 differentiation 3:442

- automatic... (continued)
 mesh moving techniques 3:554
 remeshing 2:362–363
 automotive headlamp panels 2:496, 2:497
 average internal work 2:446–449
 average strain 2:410, 2:446
 average stress 2:410, 2:446
 averaged Navier–Stokes equations 3:193
 averaging error estimators 1:93–97, 2:32–33, 2:52–53
 aviation *see* aerodynamics
 AVS development environment 1:544–545
 axial thrusts 2:559–560, 2:561
 axially loaded bars 2:358–359
 axisymmetric...
 elasticity 2:15–16
 flanged components 1:431–432
 necking 2:651–652, 2:653–654
 piercing 2:497, 2:498
 soil layers 2:581–582
- B-bar methods 2:122–123, 2:476
 B-differential equation systems 2:220
 B-FOIST *see* Beyond-Foist
 B-spline wavelets 1:480–481
 Babuška paradox 1:220
 Babuška–Brezzi (BB) condition
 contact discretizations 2:207
 linearized elasticity 2:22
 partial differential equations 1:295
 Signorini-type interfaces 1:400–401
 space discretization 3:161–162
 symmetric BEM/FEM 1:381–382
see also inf-condition
 backfill soil 2:558–561
 backscatter 3:285
 backward...
 Euler method 1:676, 1:677, 1:715–717, 3:369
 extrusion 2:500–502
 finite differencing 1:36–37, 1:50, 1:51
 rescaling 3:293
 time, central space finite differencing 1:51
 balance equations
 combustion 3:505, 3:520, 3:521
 elasticity 2:10–13, 2:229–231
 material responses 2:424
 turbulent flame combustion 3:520, 3:521
 viscoelastic fluid flows 3:482
 viscoplastic deformations 2:229–231
 balancing Neumann–Neumann preconditioners 1:640–641
 balloon angioplasty 2:617, 2:618
 bar scales 2:529–530, 2:531–533
 barrier functions 1:11, 1:15–17
 barrier method 2:202, 2:218
 bars
 linear elastic 1:128–130
 reliability methods 2:676–677
 tensile 2:345–346, 2:347
 uniaxial tension 2:344–345, 2:346–349, 2:351
 base vectors 2:8–9, 2:65, 2:91
 basis functions 3:108
 basis transform approximations 1:605–606
 BB *see* Babuška–Brezzi
 BCF *see* Brownian configuration fields
 BDM *see* Brezzi–Douglas–Marini
- beams
 Bernoulli cantilever 2:20–21
 double cantilever 2:367–368
 elastodynamics 2:767–768
 membrane locking 2:123–124
 shakedown 2:393
 three-point bending 2:364–365, 2:366–367, 2:369
 bearing capacities 2:550–551
 Bell elements 1:78
 BEM *see* boundary element method
 bending
 arterial walls 2:609–613, 2:614
 displacements 1:202, 1:206–207
 dominated action 2:68–70
 energy 1:209, 1:215
 four-point 2:39–40
 generators 1:204–205, 1:206–207
 moments 2:559–560, 2:561
 operators 1:214
 three-point 2:364–365, 2:366–367, 2:369
 Berea sandstone soil consolidation 2:583–584
 Bernoulli cantilever beams 2:20–21
 Bernstein estimates 1:164, 1:167
 Bernstein polynomials 2:214
 Besov space 1:188, 1:192
 Bessel's differential equation 2:704
 best N -term approximation 1:188–190
 Betti–Somigliana representation 1:341, 1:344
 Bettis–Burnett unconjugated test function 2:709
 Beyond-Foist (B-FOIST) 3:559
 Bézier polynomials 2:213–214
 BGT operators 2:706
 Bi-CG algorithms 1:559–560
 Bi-CGSTAB algorithms 1:559–560, 2:707
 Bi-Lanczos method 1:573–574
 BIE *see* boundary integral equations
 Bierhuselburg tunnel 2:528–529
 bifurcation 1:6
 blood flow 3:537–538
 buckling 2:142, 2:145–152, 2:156–164
 finite element models 2:617, 2:618
 parameterized systems 1:669–673
 turbulence closure 3:314
 biharmonic operators 1:34
 biological tissue 2:605–629
 biomechanics 2:605–629, 3:527–540
 biorthogonal wavelets 1:162–163
 Biot stress tensors 2:10
 Biot–Savart integrals 3:131–132, 3:141
 birefringence patterns 3:487–488
 bisection algorithms 1:101–102
 blankholders 2:495
 blanking 2:504–506
 blast waves 3:115, 3:116
 Blatt–Ko materials 2:14
 blending
 functions 3:305, 3:309, 1:124–125
 meshfree methods 1:303–306
 p -finite element method 1:125–126
 block...
 cluster trees 1:608
 deformability 1:524–529, 1:534–535
 Gauss–Seidel smoothers 3:178–179
 Jacobi preconditioners 1:567
 partitioning 1:608–609
 preconditioners 1:566–567

- blood flow 3:3, 3:527–540
 lumped parameter models 3:530–531
 pressure 3:528–540
 three-dimensional equations 3:533–540
 velocity 3:528–540
 wave propagation 3:531–533
 body forces 2:726, 3:319–320
 body-fitted meshes 3:559
 Boeing 747
 moving boundaries meshing 1:517, 1:519
 potential flow methods 3:336–337
 wing redesign 3:390–392, 3:393, 3:396
 Bogner–Fox–Schmitt elements 1:78–80
 bond slip 2:531
 bond stresses 2:530–531
 bonded particle assemblies 1:331
 bonding 2:439, 2:554
 Boolean values/operations 1:476–477, 1:483–484, 1:492
 boron 2:417–421
 bound limit theorems 2:549–551
 boundary conditions
 acoustic field equations 2:697
 advective-diffusive equations 3:33–34, 3:38–40
 arbitrary Lagrangian–Eulerian 1:423–425
 composite laminates 2:441–445, 2:447–449
 contact mechanics 2:196
 continuous blending meshfree methods 1:505–506
 Dischler–Neumann 2:696, 2:703–708, 2:711
 elliptic partial differential equations 1:15
 essential 1:293–298, 1:299, 1:305–306
 fluid flow discretization 3:361–362
 large eddy simulations 3:291–293, 3:296
 Maxwell equations 1:726
 multibody contacts 1:321–324
 Navier–Stokes equations 3:209–210, 3:220–221
 partial differential equations 1:293–298, 1:299
 plates 1:207
 RKDG 3:110–115, 3:116, 3:117
 turbulence closure 3:305–306
 boundary element method (BEM) 1:4, 1:339–371, 1:375–409
 boundary integral equations 1:346–347
 constrained shapes 2:381
 domain 2:735–740
 dual reciprocity 2:758–760, 2:761–762, 2:763
 elasticity 2:723–731
 flows with solid boundaries 3:145–149
 fracture mechanics 2:381, 2:732–735
 geomechanics 2:545–546, 2:547, 2:564
 hierarchical matrices 1:610, 1:612, 1:614–615
 multigrid methods 1:592–595
 panel clustering 1:597–600, 1:607, 1:610, 1:612, 1:614–615
 ship hydrodynamics 3:580
 Sobolev index 1:366–371
 soil consolidation 2:564
 unbounded domains 2:702
 variational formulations 1:347–358
 viscoelastic dynamic analysis 2:751–758
see also finite element method
 boundary element space discretization 1:599–600
 boundary energy 2:744
 boundary integral equations (BIE) 1:340–347
 aerodynamics 3:534–337
 elasticity 2:719–745
 Fourier transforms 1:704, 1:713–714
 Laplace transforms 1:704, 1:713–714
- matrix compression 1:175–181
 Maxwell equations 1:735
 operators 1:342–344
 plasticity 2:719–745
 time-dependent problems 1:6, 1:703–719
 time-stepping methods 1:704–705, 1:714–719
 variational formulations 1:347–358
 weak solution 1:347–358
 boundary layers 1:209–206, 1:213–214, 1:216, 1:223–224
 boundary representation schemes 1:477–485, 1:487–489
 boundary stresses 2:727
 boundary value problems (BVP) 1:339–371
 arterial walls 2:611–612
 boundary integral equations 1:340–342
 Maxwell equations 1:734–735
 nonlinear parabolic equations 1:24–26
 bounded domains 1:510–516, 3:10–11
 bounded extension splitting 1:637–638
 Boussinesq assumption 3:421–423
 Bowyer–Watson algorithm 1:501
 Box/Top Hat filters 3:273–274
 brain 1:520–521
 branching crack topology 2:388
 branching discontinuities 1:302
 Bray–Moore–Libby analysis 3:521–523
 breaking waves 3:381
 Brenst method 1:671
 Brezzi–Douglas–Marini (BDM) elements 1:259–260, 1:261
 brick-functions 1:395
 bridge scaling 1:303
 Brownian configuration fields (BCF) 3:491, 3:493–494, 3:495–496
 Brownian dynamic simulations 3:490–491
 Brownian motion 3:137–139
 Brodyan iteration 1:657
 bubble functions
 hierarchical p -refinement 3:18–20
 Maxwell equations 1:730
 residual error estimates 1:89
 space-time 3:30
 Stokes equations 1:147
 Bubnov–Calderia weak form 1:292–293
 buckling 2:139–164
 continuation computation 2:150–164
 delamination 2:367–368
 perturbation theory 2:147 149
 shells 2:70
 stability 2:142–144, 2:145–149, 2:150–164
 buff body drag 3:189–191
 buffering 3:459
 bulk forming operations 2:496–502
 bulk moduli 2:409, 2:419
 bump naps 1:491
 Burger's equation 3:100–102
 buried structures 2:558–561
 Burnett test function 2:709–710
 business jets 3:398–399, 3:427, 3:431–433, 3:442–443
 BVP *see* boundary value problems
- cabin air conditioning 3:452–453
 CAGD *see* computer aided geometric design
 calcium-silicate-hydrates (CSH) 2:517, 2:520–521
 Calculation of Non-Newtonian Flow: Finite Elements & Stochastic Simulation Technique (CONFESSIT) 3:490–491
 camera models 1:528
 cancellation properties 1:166

cannon-blast simulations 3:115, 3:116
 canonical prolongations 1:587, 1:588–589
 canonical restrictions 1:587
 cantilever beams 2:20–21, 2:367–368
 capillaries 2:519–521, 2:524–525, 3:530
 capsule implosion simulation 3:115, 3:117
 car seats 1:517, 1:519
 Caradonna–Tung rotor 3:104
 carbonate soils 2:555–558, 2:559
 cardiovascular system 3:527–540
 carotid artery 2:613, 2:614, 3:537–538
 carpet plots 1:532–533
 Cauchy coordinates 3:64–65, 3:277–278, 1:145–146
 cascade wavelet transforms 1:161–162
 case tables 1:532, 1:533
 Cauchy...
 Green tensors 2:9–10
 integral 1:718–719
 principal value integrals 1:343
 Schwarz inequality 2:31–32, 2:33–36
 singular integral equation (SIE) 1:344–345
 stress 2:613–615, 2:617
 stress tensors 1:418, 2:66, 2:456, 1:136
 causal Green's function 3:29–30
 cavity flows 3:209, 3:223–225
 cavity shape sensitivity 2:743
 Cayley–Hamilton theorem 3:312
 CBS *see* characteristic based split methods
 CDC *see* Control Data Corporation
 CDM *see* continuum damage mechanics
 C₀'s lemma 1:75–76, 1:358–362
 CEBE *see* clustered-element-by-element
 CEI EnSight Gold application 1:546
 cells
 averages 1:171–172
 centered finite volumes 1:442, 1:467
 integration 1:293
 visualization 1:528–529, 1:530–531
 CEMA *see* Constraint Energy Momentum Algorithm
 cement paste scale 2:517–521, 2:522–524
 cemented sand 2:555–558, 2:559
 centered moving least squares 1:288–289
 central circular holes plates 2:326–327
 central differencing
 block deformability 1:326
 convection–diffusion equations 1:37
 finite differencing 1:50
 method of characteristics 1:47–50
 potential flow equation 3:339–340, 3:342–343
 time integration 2:183–184
 wave equation 1:30–31
 centrifugation 3:320
 CFD *see* computational fluid dynamics
 CFL condition 1:33, 3:247
 CPM *see* computational fracture mechanics
 CG *see* conjugate gradients
 channel flows 3:209–210, 3:212–213, 3:221–223, 3:227–228
 characteristic...
 aerodynamic schemes 3:358–359
 based split (CBS) method 2:590–591, 3:583
 bifurcations 1:669–670
 Galerkin method 3:187, 3:188
 length parameters 3:592–593
 splitting 3:357
 variables 3:109–110

charge density 1:724
 Chebyshev polynomials 1:143–144
 chemical reactions 1:680, 1:582
 Cholesky decompositions 1:554–555, 1:562, 1:565
 Chorin projection scheme 3:168–170
 Christoffel symbols 1:212
 civil aircraft 3:431–433, 3:434
 clamped...
 elliptic shells 1:212–215
 hemispherical shells 2:69–70
 plates 1:224, 2:122
 spherical shells 1:224–225, 1:228
 clamping 2:617
 classical
 aerodynamics 3:325–326
 boundary integral equations 1:175–176
 laminate plate theory (CLPT) 2:453–454
 plasticity 2:552
 shakedown 2:296–299
 classifications
 convergence 1:126–127
 shells 1:215
 Sobolev index 1:366–367
 Clausius–Duhem inequality 2:12–13, 2:466
 Clausius–Planck inequality 2:13, 2:272
 clearance 2:505–506
 Clement operator 1:59–60, 1:66–68
 clipping 1:540
 closed-die rail forging 2:497–500
 closed-form constitutive equations 3:490–491
 closest-point-projection (CPPM) equations 2:228–229, 2:244–246, 2:249–259, 2:262–263
 closures
 filtered Navier–Stokes equation 3:32–33
 mesh refining 1:102, 1:103
 Smagorinsky eddy viscosity 3:43–44
 turbulence 3:301–322
 clouds 1:291–292
 CLPT *see* classical laminate plate theory
 clusters
 bifurcations 2:160
 clustered-element-by-element (CEBE) 3:566
 hierarchical matrices 1:608
 multipole expansions 3:142–144
 panel clustering 1:5–6, 1:597–615, 2:731
 trees 1:601–602, 1:604, 1:606, 1:608
 coarse grids
 corrections 1:578, 1:582, 1:584–585, 1:588–589
 dynamic multilevel methods 3:233
 matrices 1:588
 scales 3:11–27
 coarsening
 adaptive mesh design 1:104
 adaptive wavelets 1:193
 flow field compression 1:170
 local mesh refinement 1:98, 1:104
 coated steel sheets 2:474–476
 coaxial supersonic jets 3:447
 code interfaces 3:428
 coefficients
 drag 3:101–103, 3:122–123, 3:146–147, 3:174–175
 force 3:373
 Fourier 3:243–244
 lift 3:84–85, 3:146–147
 positive 1:456–457, 1:463–464, 3:349–356

predictions 1:189, 1:191–193
 pressure 3:101, 3:103, 3:435
 of variation (COV) 2:567–568, 2:662–663
 coercivity 3:33
 cohesive crack propagation 2:379
 cohesive-frictional soils 2:550–551
 cohesive-zone models 2:337, 2:349–355, 2:363–367
 coining process 1:431, 1:433
 cold jets 3:446–447
 collapse 2:150, 2:160, 2:163, 2:537–538
 collapsed Cartesian coordinates 3:64–65, 1:145–146
 collocation
 algebraic polynomial expansions 1:143–145
 boundary element method 2:723–727, 2:732–733
 boundary integral equations 1:346, 1:704, 1:712–713
 discretization 1:292, 1:660
 fracture mechanics 2:732–733
 partial differential equations 1:292
 space-time boundary integrals 1:712–713
 time integration 2:184
 color mapping 1:531–532, 1:533–534
 combustion 3:3, 3:499–523
 governing equations 3:504–507
 instable flames 3:500–501, 3:503–504
 laminar flames 3:500–501
 mixture fractions 3:508–511
 nonpremixed flames 3:500–501, 3:502
 overall reaction terminology 3:507–508
 partially premixed flames 3:500–501, 3:503
 premixed flames 3:500–502
 reaction terms 3:507
 regimes 3:500–504
 stable flames 3:500–501, 3:503–504
 stoichiometry 3:507–508
 turbulent flames 3:500–501, 3:514–523
 compact-tension (CT) 2:361–362
 compaction 1:431–433
 compatibility conditions 3:118, 3:468–471
 complementary...
 Ditchfield energy principle 2:24
 energy functions 2:251
 operators 1:229
 partitioning 2:422
 completely discretized schemes 1:31–32
 complex...
 geometries 3:61–88, 3:346–348
 multidimensional domains 3:359–365
 shapes 1:475–494
 viscoelastic fluid flows 3:489–490
 complexity
 adaptive wavelets 1:187–195
 panel clustering techniques 1:603–604
 composite laminates 2:431–458
see also fiber-reinforced...
 composite matrices 1:256–257
 compound bifurcation points 2:159–164
 compressed foam-rubber tubes 2:53–54
 compressed matrices 1:175–181
 compressible...
 Euler equations 3:101, 3:103–105, 3:383–386
 flows 3:2–3
 discontinuous Galerkin methods 3:74–76, 3:77–81, 3:122–123
 moving domains 3:76–77
 plasma 3:81–85
 shock capturing 3:551–553

stabilization parameters 3:551–553
 vortex particle methods 3:149–150
see also turbulence...
 hyperelastic materials 2:13–14
 Navier–Stokes equations 3:122–123
 partition solution procedures 2:582
 strain-energy functions 2:13–16
 turbulence closure 3:302
 compression
 failure 2:438
 flow fields 1:169–175
 geomechanics 2:555–556
 compressive loadings 2:352, 2:354–355, 2:535–536
 computability
 computational fluid dynamics 3:2, 3:183–205
 direct numerical simulations 3:184–185, 3:191–197
 large eddy simulations 3:184–185, 3:191–197
 computational...
 aerodynamics
 Euler equations 3:327, 3:329
 Navier–Stokes equations 3:327, 3:329
 potential flow methods 3:334–348
 transonic flow 3:326–327
 aeroelasticity 3:459–477
 biomechanics 2:605–629
 complexity 3:87–88
 contact mechanics 2:195–221
 costs 3:389–390, 1:132
 discrete element code 1:316
 flow mechanics
 aeroacoustics 3:444–448
 aerodynamics 3:407–454
 code interfaces 3:428
 Euler codes 3:413–416
 flutter analysis 3:448–449, 3:450
 fundamental studies 3:436–437
 historical overviews 3:408–413
 large eddy simulations 3:423–426
 mesh generation 3:426–428
 Navier–Stokes code 3:416–418
 numerical codes 3:413–418
 paraflois 3:449–451
 research 3:436–437
 Reynolds averaged turbulence 3:418–423
 shape optimization 3:437–443
 validation 3:428
 Dassault Aviation 3:407–454
 military aircraft 3:428–431
 fluid dynamics (CFD)
 adaptivity 3:2, 3:183–205
 computability 3:2, 3:183–205
 discontinuous Galerkin methods 3:91–123
 mesh generation/adaptivity 1:497–521
 nonlinear aeroelasticity 3:459–477
 RKDG 3:97, 3:104–115, 3:116, 3:117
 ship hydrodynamics 3:580
 transonic flow 3:327
 turbulence closure 3:301–322
 forming processes modeling 2:461–507
 fracture mechanics (CFM) 2:375–402
 cracking process 2:376–377
 geometrical representations 2:394–400
 nongeometric representations 2:394–400
 geomechanics 2:543–569
 goal-oriented error estimators 1:98
 grids 1:547–548

computational (*continued*)
 homogenization 2:413–414
 interfacing visualization systems 1:546–548
 nonlinear aeroelasticity 3:459–477
 rheology 3:481, 3:487–488, 3:489
 visualization 1:5, 1:525–548
 computer aided geometric design (CAGD) 1:667
 concrete 2:513–539
 autogenous shrinkage 2:514, 2:516–529
 cooling tower analysis 2:529–538
 embedded anchor bolts 2:360–361
 tension stiffening 2:529–538
 under fire 2:576, 2:598–599
 condition numbers 1:365–366
 Galerkin methods 1:365–366, 3:119–120
 local convergence theory 1:655–656
 Signorini-type interfaces methods 1:398–399
 thin-walled structures 2:125
 CONFESSITI *see* Calculation of Non-Newtonian Flow: Finite Elements & Stochastic Simulation Technique
 configuration field evolution 3:493
 conforming finite elements 1:73, 1:77–85
 conforming Galerkin method 1:177
 conjugate gradients (CG)
 acceleration 1:592
 direct linear algebraic solvers 1:557–558
 material responses 2:416–417
 nonlinear parabolic equations 1:26
 preconditioned 1:642–643, 1:147
 conjugated test functions 2:709–710
 connectivity, visualization 1:541
 conservation
 of energy 1:30, 1:419–420, 1:426, 2:11–13
 equations
 advective-diffusive 3:33, 3:34
 arbitrary Lagrangian-Eulerian 1:419–420
 exact potential flow 3:343–345
 fluid flow 3:330–331
 reacting flows 3:505, 3:514
 Godunov finite volume discretizations 1:443
 laws
 adaptive wavelets 1:169
 arbitrary Lagrangian-Eulerian 1:424–425
 finite volumes 1:439–450, 1:464–470
 geometric 1:424–425, 3:466–468
 nonlinear 1:439–450, 1:468–470, 3:491, 3:596–115, 1:150
 scalar 1:439–450, 3:97–98, 3:100–108, 3:349–351
 shock capturing 3:356
 spectral methods 1:148–150
 of linear momentum 2:11
 of mass 2:10
 time integration 2:185–187
 conservative, . . .
 derivatives 3:129
 difference equations 3:341–342
 discontinuous Galerkin methods 3:92, 3:119
 consistency
 advective-diffusive equations 3:34, 3:35, 3:36
 continuous moving least squares 1:286
 discontinuous Galerkin methods 3:119
 errors 1:106
 Godunov finite volume discretizations 1:443
 moving least squares 1:286, 1:289
 small strain elastoplasticity 2:738

Sobolev index 1:367–368
 window functions 1:283
 consistent integration, plastic flow 2:274–275
 consistent tangent operators (CTO) 2:736–738
 consolidation, geomechanics 2:543, 2:561–567
 constant mass density 2:684–685
 constant total enthalpy 3:358–359
 constitutive, . . .
 behavior, composite laminates 2:449–451
 equations
 arbitrary Lagrangian-Eulerian 1:428–429
 inelastic materials 2:638–640
 inverse 2:640–641
 material parameter identification 2:637–634
 shell theory 2:95–98
 viscoelastic direct boundary elements 2:752–753
 viscoelastic fluid flows 3:481–482, 3:490–491, 3:493–494
 integration 2:736–740
 laws 2:310, 2:204–206, 2:621, 2:624
 models
 arterial wall mechanics 2:607–609
 geometricals 2:551–558, 2:559
 heart wall mechanics 2:619–622
 ligaments 2:626–629
 material frame indifference 2:239–240, 2:241
 single crystal plasticity 2:273
 nongeometric representations 2:394–399
 tensors 2:13–14
 constrained Delaunay kernel 1:508
 constrained shape methods 2:378–394
 constraint, . . .
 equations 2:200
 impositions 1:321–324
 solution methods 2:197
 Constraint Energy Momentum Algorithm (CEMA) 2:186–187
 constructive solid geometry (CSG) 1:483–485
 contact, . . .
 conditions regularization 1:314–316
 constraint impositions 1:321–324
 detection 1:317–321, 2:215
 discontinuity approximation 3:111, 3:112
 discretizations 2:206–214
 force evaluation 1:516, 1:517, 1:521–324
 friction modeling 2:471–476
 mechanics 2:195–221
 algorithms 2:197, 2:198, 2:214–221
 arbitrary Lagrangian-Eulerian 1:431, 1:433
 contact discretizations 2:206–214
 continuum description 2:198–206
 detection 1:317–321, 2:215–216
 plane orientation 1:322–323
 resolution phase 1:318–321
 segments 2:209–210
 continuation principle 2:155
 continuity
 bifurcations 1:670–671
 buckling 2:150–164
 equations 1:726, 3:302, 3:505
 meshfree methods 1:287, 1:288, 1:289
 parameterized nonlinear equations 1:664–666
 projection-based interpolation 1:733
 continuous, . . .
 blending 1:303–306
 boundary integral operators 1:351
 conservation laws 1:424–425

discretization 2:664–665
 interpolation 3:484
 moving least squares 1:285–286
 pressure interpolation 1:262–264
 yielding plasticity 2:552–553
 continuum, . . .
 constitutive modeling 2:462–469
 contact mechanics 2:198–206
 damage mechanics (CDM) 2:353–362, 2:440–441, 2:452–458
 failure modeling 2:335–336
 mesh sensitivity 2:341–349
 partition-of-unity 2:369
 mechanics
 arbitrary Lagrangian-Eulerian 1:413–433
 Eulerian algorithms 1:413–414
 Lagrangian algorithms 1:413–414
 thin-walled structures 2:64–68
 micromechanics 2:515–519
 slip theory 2:267–268, 2:269 274
 contour dynamics 3:131, 3:132–133
 contours of vorticity curl 3:131, 3:132–133
 contraction flow 3:487–488
 contraction principle 1:441, 1:447
 contravariant base vectors 2:65
 Control Data Corporation (CDC) 3:327
 control theory 3:381–383
 control volumes 1:458, 1:460–461
 convection
 arbitrary Lagrangian-Eulerian 1:429, 1:430–431
 convection-diffusion
 discontinuous Galerkin methods 3:120–123
 equations 1:35–50, 1:696–700, 3:149
 turbulence 3:196
 convective gradient projections 3:586
 convective velocity 1:417
 discontinuous Galerkin methods 3:91, 3:115, 3:120–123, 3:196
 shallow water equations 3:242
 Conventional Serial Saggered (CSS) time integrators 3:471–473
 convergence
 acceleration 3:339, 3:365–379
 adaptive finite element methods 1:98, 1:103
 advective-diffusive equations 3:36
 buckling computation 2:154–155
 characteristics method 1:126–131
 conservation laws 1:446–450, 1:467–468, 3:97, 3:98–104
 contact mechanics 2:197–198, 2:219
 Dirichlet-to-Neumann boundary condition 2:707
 discontinuous Galerkin methods 3:96, 3:97
 domain decomposition 1:627–630
 eigenvalues 1:570
 factors 1:652–653
 Koiter shell theory 1:217
 meshfree methods 1:287
 multigrid methods 1:583–584
 optimality orders 1:362–364
 orders 1:652–653
 p -finite element method 1:126–131
 rate optimality 1:207
 rates 1:185–187
 RK4DO 3:110
 shakedown 2:316
 shells 1:214–215
 Sobolev index 1:367–368
 symmetric BEM/FEM 1:380–382
 theory 1:649–669

converse piezoelectric effect 2:761
 convolution quadrature 2:754, 2:756–757
 cooling tower analysis 2:529–538
 coordinates
 Cartesian 3:64–65, 3:277–278, 1:145–146
 mesh filtering 3:278
 plates 1:199–200, 1:202
 polynomial expansions 3:64–65, 3:66
 shells 1:199–200
 transformations 1:535
 copper crystals 2:283–284, 2:286–287
 Coriolis terms 3:242, 3:251
 covek of separation 2:350–351
 corner singularities 3:157 158
 corner-to-corner contacts resolution 1:318, 1:320–321, 1:322–323
 corrected smooth particle hydrodynamics (CSPH) 1:291
 correctors 2:153–156, 3:414
 corrosive wear 2:206
 Cosserat continuum 2:355–356
 Cosserat surfaces 2:76
 cost functions 3:439–440
 costs
 aerodynamic shape optimization 3:389–390
 structural dynamics 2:188–189
 Coulomb friction 1:315, 1:328, 1:433, 2:205–206
 coupled
 boundary element/finite elements 1:375–409, 2:21–23, 2:547
 elastoplasticity 2:744–745
 fast solvers 1:389–394
 geomechanics 2:547
 least squares 1:394–396
 Signorini-type interfaces 1:376–377, 1:396–403
 symmetric coupling 1:377–389
 variational formulation 1:398–403, 1:405–408
 damage-plasticity models 2:340–341
 direct numerical simulations 3:184–185, 3:191–198
 discretization error estimation 2:40–46
 finite elements/boundary elements 1:375–409, 2:21–23, 2:547, 2:744–745
 finite elements/discrete elements 1:325–326, 1:334–335
 finite elements/meshfree methods 1:293, 1:303–306
 fluid/fluid-mesh/structure time integrators 3:471–473
 free-surface equations 3:580
 gradient plasticity-damage models 2:359–360
 large eddy simulation 3:184–185, 3:191–198
 nonlinear aeroelasticity 3:459–466
 overlapping domains 2:575–599
 a posteriori error estimates 2:40–46
see also dual formulations
 coarse staggered meshes 3:223–225
 COV *see* coefficient of variation
 covariance
 base vectors 2:65, 2:91
 constitutive equations 2:646, 2:647
 derivatives 1:211–212
 random variables 2:659
 covering admissibility 1:601–602, 1:606
 crack, . . .
 closure effect 2:483
 concrete mechanics 2:535–538
 discontinuous meshfree methods 1:300–303
 growth
 arbitrary 2:382–394
 boundary element method 2:734–735
 conceptual models 2:377

crack... (continued)
 constrained shape methods 2:378–394
 resistance 2:396–399
 simulation 2:377
 inelastic bodies 2:311
 meshfree methods 1:280
 mouth sliding displacements 2:397
 opening modes 2:350–351
 plates 2:39–40
 process representation 2:376–377
 propagation 2:352, 2:354–355, 2:366–369
 shape sensitivity 2:743
 cranium bone 1:520–521
 Crank–Nicolson finite differencing 1:51
 creep 2:526
 crew rescue vehicles (CRW) 3:434–435
 crew transfer vehicles (CTV) 3:434–435
 critical...
 conditions 2:145
 loads 2:159–160
 points 1:453, 1:536–537, 2:156–158
 slip resistance 2:277
 states 2:139, 2:144 149, 2:156–159, 2:161–164
 cross...
 constraint method 2:202
 diffusion 3:309
 flows 1:425–426, 3:487–488
 stress tensors 3:274
 Crouzeix–Raviart elements 1:264–266, 1:270–272, 1:465–466, 1:589
 cruise missiles 3:429, 3:430
 CRW *see* crew rescue vehicles
 cryogenic wind-tunnels 3:431
 crystal lattices 2:339–340
 crystal plasticity 2:267–287
 continuum slip theory 2:267, 2:269–274
 elastic–plastic tangent moduli 2:279–280
 free energy 2:271–272
 heterogeneous polycrystallines 2:281, 2:282–283, 2:286–287
 homogenization 2:269–271, 2:281–282, 2:286–287
 numerical examples 2:283 287
 polycrystallines 2:267–287
 shearing 2:283–284
 updating 2:274–280
 variational formulations 2:280–283
 crystal sheets 2:284–286
 crystal strips 2:284
 CSG *see* constructive solid geometry
 CSH *see* calcium-silicate-hydrates
 CSIE *see* Cauchy singular integral equation
 CSPH *see* corrected smooth particle hydrodynamics
 CSS *see* Conventional Serial Staggered
 CT *see* compact-tension
 CTO *see* consistent tangent operators
 CTV *see* crew transfer vehicles
 cube within water 3:604–607
 cubic spline window functions 1:281, 1:283–284
 culling 1:542
 cumulative pore-size distributions 2:519
 curl operator 1:725, 1:727–734
 currents
 configuration 2:64–66, 2:76–77, 2:81
 Maxwell equations 1:726
 curvatures
 mesh adaptivity 1:515–516
 shell theory 2:93–95

tensors 1:211–212
 thickness locking 2:121, 2:124–125
 unit meshing 1:504–507
 curved...
 boundary conditions 3:114
 boundary difference approximations 1:13
 domains 1:109–111
 segment discretization 1:505
 CUSP schemes 3:358–359
 cutting, visualization 1:540
 cutting-plane algorithm 2:246
 cyclic loadings 2:626–627, 2:648–649
 CYCLONE 2:320
 cylinders
 compressible flow 3:122–123
 drag 3:191–192, 3:193–194, 3:197–198
 flow past 3:122–123, 3:201–205, 3:487–488, 3:494–496
 laminar flow 3:201–205
 plasma flow 3:84–85
 cylindrical...
 billet upsetting 2:479
 heat sources 2:588–589
 shells
 bending 2:84–102
 buckling 2:163
 eigen-frequencies 1:224, 1:226–229
 inflation 2:84–102
 intersections 1:133
 membrane locking 2:124, 2:125
 window functions 1:282

D/BEM *see* domain boundary element method

Dahlquist barrier 2:178
 damage
 composite laminates 2:431–458
 concrete under fire 2:598
 energy release rates 2:456–458, 2:482
 evolution 2:453
 loading functions 2:336
 mechanics 2:336–349, 2:440–441, 2:456–458
 microstructure 2:338–339
 continuum models 2:355–362
 coupled damage-plasticity 2:340–341
 discrete failure models 2:362–370
 isotropic 2:336–338
 material instabilities 2:341–349
 mesh sensitivity 2:341, 2:344–349
 microplane 2:339–340
 partition-of-unity 2:365–369
 shakedown 2:311
 surfaces 2:453, 2:456–457
 tensors 2:453–454
 thresholds solids 2:482
 variables 2:452–453
 Damböcker numbers 3:521
 damping
 buckling 2:163
 deformation analysis 1:332–333
 Jacobi iteration 1:581–583
 Jacobi smoother 3:178
 nonlinear systems 1:659–660
 raises 2:179
 shallow water equations 3:241
 viscosity 3:305–306
 dams 2:592, 2:593, 2:594

Darcy's law 2:582
 Dassault Aviation 3:407–454
 data...
 extraction 1:538–541
 fields 1:528–529, 1:530–531
 flow 1:543
 forms 1:528–531
 interpolation 1:531
 streaming 1:543
 transfer 1:547–548
 visualization 1:525–526, 1:528–531, 1:538–541, 1:543, 1:547–548
 dataset attributes, visualization 1:528–529, 1:530–531
 Daubechies, I. 1:162
 DCDD *see* discontinuity capturing directional dissipation
 DD *see* domain decomposition
 DDA *see* discontinuous deformation analysis
 DDM *see* direct differentiation method
 de Rham diagrams 1:729, 1:732–734
 decimation 1:541
 decohesion relations 2:350
 decomposition
 Cholesky 1:554–555, 1:562, 1:565
 dynamic multilevel methods 3:219–228
 Germano's 3:274–275
 hierarchical two-level 1:385–386
 large-scale 3:219–228
 Leonard's 3:274
 material responses 3:425
 nonoverlapping domains 1:91, 1:618–620, 1:633–644, 2:712
 overlapping domains 1:617–620, 1:630–633
 triple 3:293
see also domain...
 deconvolution subgrid-scale modeling 3:287–288
 deep drawing 2:284–286, 2:496, 2:497
 defect correction 1:40–43, 3:20
 deformation
 arterial walls 2:608
 blankholders 2:495
 body collisions 1:314–335
 clamped elliptic shells 1:213
 deforming-spatial-domain stabilized spacetime (DSD/SST) 3:549, 3:555–557, 3:571–574
 elastic bodies 2:7–9
 fields method (DFM) 3:491, 3:492–493, 3:495–496
 gradients 2:8–9, 2:56, 2:463, 2:477–478
 medial modeling 1:486, 1:487, 1:489–490
 mesh adaptivity 1:514–515
 moving boundaries 1:519
 plates 1:203–205
 porous medium 2:584–586
 shell theory 2:76–77, 2:81
 thin-walled structures 2:65, 2:76–77, 2:81
 degenerating solid elements 2:79–83
 degradation, geomechanics 2:553
 degrees of freedom (DOF)
 material responses 2:416–417
 mixed finite element methods 1:238, 1:258–261, 1:263–266
 Stokes equations 1:263–266
 thermal diffusion 1:258–261
 delamination
 buckling 2:368
 cohesive-zone models 2:363
 composite laminates 2:453
 partition-of-unity concept 2:367–368
 Delaunay methods 1:498, 1:501–502, 1:508

DEM *see* diffuse element method; discrete element methods
 density
 airfoils 3:101, 3:103
 constant mass 2:684–685
 free charge 1:724
 mesh 2:417, 2:418
 normalized normal 2:659–660
 probability density function 2:567–568, 3:520–521
 profiles 1:431, 1:432–433
 Deny-Lions lemma 1:60–61, 1:67
 dependent stress 2:608
 derivatives
 conservative 3:129
 constitutive equations 2:650–651
 covariance 1:211–212
 diffuse 1:289–290
 elliptic partial differential equations 1:14
 fractional time 2:752–753
 Fréchet 1:182
 Lie 2:243
 local time 1:418
 material 1:418–419, 3:129
 normal derivative kernels 1:605
 pseudospectral 1:142
 shape functions 1:283–291
see also time...
 derived statistics 2:670–671, 2:675–676
 design
 aerodynamic shape optimization 3:379–381, 3:389–390
 costs 3:389–390
 stochastic finite elements 2:677, 2:678, 2:679–680
 destructuring, geomechanics 2:554–555
 detachment points 1:536
 detection 1:317–321, 1:535, 2:196, 2:214–216
 deterministic eddies 3:321
 deterministic geotechnical analysis 2:544–545
 development environments, visualization systems 1:543, 1:544–545
 deviatoric stresses 3:593–594
 DEVSS *see* discrete elastic-viscous stress splitting
 DFM *see* deformation fields method
 DG *see* discontinuous Galerkin
 DGCL *see* discrete geometric conservation laws
 DIA *see* direct interaction approximation
 differencing
 aerodynamics 3:340–342
 boundary values 1:23–25
 convection–diffusion equations 1:44–45
 curved boundary approximations 1:13
 elliptic partial differential equations 1:12–13
 equations 3:340–342
 five-point 1:9, 1:50
 meshes 1:45–47, 1:290, 1:291
 Murman–Cole 3:339–340, 3:341–343
 nine-point approximations 1:13–14, 1:50
 potential flow equation 3:339–340, 3:342–344
 transonic small-disturbance equation 3:340–342
 two-point boundary problems 1:10–11
 vectors 2:80, 2:81
 wave equation 1:30–34
see also central differencing; finite difference
 differential...
 equations
 Bessel's 2:704
 conservation 1:419
 mixed finite element methods 1:238

differential... (continued)
 parabolic 1:675–702, 3:341–342
see also ordinary...; partial...
 geometry 1:493–494, 2:63–64
 operators 1:238, 3:30–32
 quadrature methods 2:184

differentiation...
 continuation methods 1:664–665
 error 2:741, 3:282
 shape perturbations 2:740–741
 diffuse derivatives 1:289–290

diffuse element method (DEM) 1:280, 1:289–290

diffusion
 advective-diffusion equation 3:25–26
 conservation laws 1:466–468
 dominated convection–diffusion–reactions 3:196
 equation 1:18–28, 1:696–700, 3:25–26
 flames 3:500–501, 3:502
 incompressible viscous flows 3:170–171
 reaction equations 1:696–699, 1:700
 soil consolidation 2:562–563
 streamlines 1:449–450, 3:97, 3:187, 3:188–189
 thermal 1:238–240, 1:252–254, 1:257–262, 3:503
 turbulence closure 3:304, 3:309, 3:319–320
 viscous flows 3:137–139
see also convection–diffusion

diffusion 1:289, 2:557, 2:558

dilute family methods 2:412

dimensional reduction 2:70–107

direct constraint elimination 2:203

direct differentiation method (DDM) 2:667, 2:741–742

direct interaction approximation (DIA) 3:51

direct methods
 acoustics 2:713
 buckling stability analysis 2:157
 eddy viscosity 3:51
 hierarchical matrices 1:614–615
 linear algebraic solvers 1:553–560
 nonlinear parabolic equations 1:27
 panel clustering 1:599
 shell theory 2:76–79, 2:83

direct numerical simulations (DNS)
 adaptive 3:184–185, 3:191–197
 applications 3:295–294
 computability 3:184–185, 3:191–197
 eddy viscosity 3:49, 3:51–52
 forced homogeneous turbulence 3:236
 large eddy simulations (DNS/LES) 3:184–185, 3:191–198
 Navier–Stokes equations, turbulence 3:218–219
 numerical error 3:281–282
 reacting flow combustion 3:506
 renormalized scales 3:261
 resolution requirements 3:279–281
 shallow water equations 3:247
 time advancement schemes 3:283
 turbulence 3:2, 3:236, 3:269–270, 3:279–283, 3:293–294
 turbulent flames 3:514–515, 3:518

direct piezoelectric effect 2:761

director...
 field interpolation 2:112
 spaces 1:207
 vectors 2:118

Dirichlet boundary conditions
 continuous blending meshfree methods 1:305–306
 error estimates 2:36–37

far field scales 3:9

shallow water equations 3:245–246, 3:248–249, 3:250, 3:251

thin-walled structures 2:67

variational multiscale method 3:11–18

Dirichlet principle 1:348

Dirichlet problems
 boundary integral equations 1:707, 1:714–715
 hierarchical error estimators 2:52
 implicit residual error estimators 2:30–31
 panel clustering 1:599
 preconditioners 1:638

Dirichlet-to-Neumann (DtN) boundary conditions
 acoustics 2:696, 2:703–708, 2:711
 multiscale methods 3:8–11
 stabilized methods 3:8–11

discontinuity capturing directional dissipation (DCDD) 3:550–551

discontinuity failure modeling 2:337

discontinuous...
 deformation analysis (DDA) 1:326–329, 1:330–331, 1:332–333
 finite elements 3:91–123
 Galerkin (DG) methods 3:1–2
 advective-diffusive equations 3:36–37
 complex viscoelastic fluid flows 3:491, 3:495
 compressible flows 3:74–76, 3:77–81
 computational fluid dynamics 3:91–123
 conservation laws 3:91, 3:96–115
 convection 3:91, 3:115, 3:120–123
 diffusion 3:120–123
 enhancement 3:93–94
 hp-finite methods 3:62, 3:71–75, 3:77–81
 linear hyperbolic equations 3:91, 3:92–96
 of order zero 1:677, 1:683–684
 Runge–Kutta 3:97, 3:104–117
 second-order ellipticity 3:91, 3:115–120
 viscoelastic fluid flows 3:483, 3:484, 3:495
see also Petrov–Galerkin

interpolation 1:264–266, 3:484

meshfree methods 1:300–303

modeling 1:311–335, 2:537, 2:569

pressure interpolation 1:264–266

discontinuum, damage models 2:369

discrete...
 bounded extensions 1:638
 contact-friction 2:473–476
 crack models 2:362
 curve discretization 1:505
 elastic shakedown 2:301–304
 elastic-viscous stress splitting (DEVSS) 3:483–487, 3:495
 element methods (DEM) 1:3–4, 1:311–335
 arbitrary shape fracture mechanics 2:390–394
 basic framework 1:314–316
 block deformability modeling 1:324–329, 1:334–335
 boundary conditions 1:321–324
 contact constraint impositions 1:321–324
 contact detection 1:317–321
 discontinuous deformations 1:326–329
 energy balance contact 1:331–333
 fracturing 1:329–331, 2:390–394
 fragmentation 1:329–331
 geomaterials 2:545–546, 2:548
 interacting bodies 1:317–321
 nonsmooth contact conditions 1:314–316
 temporal discretizations 1:331–333
 time integration 1:331–333

estimate 1:82

failure modeling 2:335, 2:362–370

finite volume methods 1:443–444

function representation 1:318, 1:320–321

geometric conservation laws (DGCL) 1:425, 3:466–468

interfaces 1:424

iterative solutions 1:376

Kirchhoff elements 2:110

maximum principle 1:10–11, 1:445–446

Morse theory 1:493

moving least squares 1:286–287

Navier–Stokes equations 3:219–228

optimality conditions 2:301

Poincaré inequality 1:82

projection smoothers 3:179

Sobolev inequality 1:82

test filters 3:279

unit meshing surfaces 1:506

discretization 1:3–4
 boundary integral equation method 1:715
 collocation boundary element method 2:723–724
 contact mechanics 2:206–214
 continuous blending meshfree methods 1:304
 curved segment 1:505
 discrete element methods 1:331–333
 dynamic equations 2:492–493
 dynamic multilevel methods 3:231–240
 enhanced 3:560–561, 3:566–567, 3:568–570
 error control 2:37–39, 2:40–46
 error estimation
 convection–diffusion equations 1:58–60, 1:46–47, 1:49–50
 coupled methods 2:40–46
 elliptic partial differential equations 1:15–17
 model adaptivity 2:44–46
 a posteriori 2:40–46
 0-scheme 1:22–23
 wave equation 1:32–33

finite difference 1:3, 3:360–361

finite volume 1:465–466, 3:360–361

fluid dynamics 3:560–561, 3:566–567

fluid flows 3:359–365

Galerkin 1:600, 1:606, 1:607, 1:612, 3:170

General Galerkin 2:700–701, 3:184–185, 3:187–189, 3:191–198

geomaterials 2:544–545

Helmholtz equation 2:696, 2:698–702

homogenization 2:281–282, 2:286–287

interface-capturing technique 3:560–561

material responses 2:416–417

mesh refinement 3:560–561

multigrid methods 1:588–589

Newton's method 1:658

nonlinear aeroelasticity 3:462–464

panel clustering 1:599–600

partial differential equations 1:291–300

plates 1:201, 1:220–221

shells 1:201, 1:220–221

stochastic finite elements 2:663–667

thin-walled structures 2:112–113

turbulence closure 3:219

unit meshing 1:504–505

up-dating 3:568–570
see also finite element...; spatial...; time

dispersion
 error 2:696, 2:700, 2:701–702
 numbers 1:33–34
 time integration 2:179

displacement
 boundary conditions 2:311, 2:442–443, 2:447, 2:449
 boundary integral equations 1:345, 2:721–722
 composite laminates 2:442–443, 2:447, 2:449
 compressed foam-rubber tubes 2:53–54
 discontinuity 2:722–734

fields
 concrete mechanics 2:525–526
 fracture mechanics 2:399–400
 hydroelastic cloaking 2:689–691
 layerwise laminate theory 2:435–436
 structural-acoustics 2:684–687

gradients 1:244–245, 2:343–344

integral representation 2:720–721

magnitude 2:85, 2:87–91

mapping 1:533–534, 2:247, 2:490–491

parameterization 2:116–120

plates 1:200–207, 1:209–210, 1:215, 1:229

potential 2:685–686, 2:690

representative volume element 2:442–443

retare-mapping 2:247

shakedown 2:511

shells 1:200–201, 1:215

statically determinate structures 2:662–663

symmetric gradients 1:244–245

weighting functions 2:172–173

dissipation 2:465

contact mechanics 2:206
 crystal plasticity 2:272–273
 eddy viscosity 3:52–53
 forced homogeneous turbulence 3:237–239
 function regularization 2:320
 heterogeneous microstructures 2:282–283
 hyperelastic 2:465
 inequality 2:465
 time integration 2:179, 2:186–187, 3:52–53
 turbulence closure 3:304, 3:308, 3:318, 3:319

wave equations 1:707

distortion-free refinement 2:38

disturbed state concept (DSC) 2:554–555

divergence-free Fourier–Legendre polynomials 3:215–217

DML *see* dynamic multilevel methods

DNA molecules 1:517, 1:519

DNS *see* direct numerical simulations

DNS/LES *see* direct numerical simulation/large eddy simulations

DOF *see* degrees of freedom

DoE–Edwards model 3:492–493, 3:495

domain...
 acoustics 2:702–703, 2:712–713
 adaptive wavelets 1:163–165
 arbitrary Lagrangian–Eulerian method 1:414, 1:416, 1:417–418
 boundary element method (D/BEM) 2:735–740
 boundary integral equations 1:704, 1:713–714
 bounded 1:510–516, 3:10–11
 complex multidimensional 3:359–365
 compressible flow 3:76–77
 curved 1:109–111
 decomposition (DD) 1:6
 acoustics 2:712–713
 adaptive wavelets 1:163–165
 finite elements 1:618–619
 historical overviews 1:619–621
 incomplete LU factorization 1:567
 material responses 2:425
 nonoverlapping 1:91, 1:618–620, 1:633–644, 2:712

- domain... (continued)
 overlapping 1:617–618, 1:619–620, 1:630–633
 preconditioning 1:617–644
 of dependence 1:29–30
 incompressible flows 3:69–70, 3:157–158
 integrals 2:726
 Lipschitz 1:167
 overlapping 1:617–618, 1:619–620, 1:630–633, 2:575–599
p-finite element method 1:151
 plates 1:199–200
 preconditioners 1:566–567
 shakedown 2:296–297, 2:298–299
 shells 1:199–200
 thin 1:199–229
 time 1:703–719, 2:713, 2:751–758
 viscoelasticity 2:754, 2:755–756
 viscous flows 3:157–158
 dominant transport 3:164–166
 double cantilever beams 2:367–368
 Doyle–Ericksen formula 2:14
 drag
 aerodynamic shape optimization 3:380
 buff bodies 3:189–191
 coefficients 3:101–103, 3:122–123, 3:146–147, 3:174–175
 correction factor 3:486–487
 dissipation 3:195–196
 laminar flow around cylinders 3:202–203
 square cylinders 3:191–192, 3:193–194, 3:197–198
 surface-mounted cubes 3:192–195, 3:196, 3:197–198
 viscoelastic fluid flows 3:495–496
 dragrace fittings 1:134–135
 drained shearing 2:557, 2:558
 drained triaxial compression 2:555–556
 drawing 2:284–286, 2:496, 2:497
 drilling rotations 2:115, 2:116–118
 drop tolerance preconditioners 1:565
 drop-by-position preconditioners 1:565
 drop-by-size preconditioners 1:565
 dropping updates 2:279
 Drunker–Prager criterion 2:535–536
 dry friction 2:205
 Dryje's preconditioner 1:635–636
 DSC *see* distributed state concept
 DSD/SST *see* deforming-spatial-domain stabilized spacetime
 DN *see* Dirichlet-to-Neumann
 Du Fort–Frankel–Soni's finite differencing 1:51
 dual formulations
 adaptive computation 1:693–694
 adaptive wavelets 1:167–168
 augmented Lagrangian 2:261–262, 2:263
 boundary elements 2:733, 2:758–763
 closest-point-projection equations 2:254–259, 2:262–263
 contact mechanics 2:218
 elastic shakedown 2:300
 error representation 1:97, 2:26–27, 2:48
 goal-oriented error estimators 1:97
 norms 1:86
 plane linear elasticity 1:377
 reciprocity 1:704, 2:758–763
 Signorini-type interfaces 1:398–403
 time-stepping 3:376–377, 3:378, 3:379
see also coupled...
 ductile...
 failure 2:535–536
 fractures 2:349, 2:350
 single crystals 2:267
 dumbbells 3:492–494
 Dunford–Cauchy representation 1:615
 DYNAD 2:212
 dynamic...
 adaptivity 1:170–174
 arbitrary Lagrangian–Eulerian mechanics 1:428
 contact 2:220–221
 equilibria stability 1:672–673
 isothermal solutions 2:589–592
 meshes 3:464–466
 multilevel methods (DML) 3:207–264
 piezoelectricity 2:758–759, 2:761–762, 2:763
 shakedown 2:310
 subgrid-scale modeling 3:288–290
 substructuring 2:686–687
 E-fluxes 1:444–445
 EALST 3:566–567
 etching crystals 2:284–286
 EARSIM *see* explicit algebraic Reynolds stress model
 eddies 3:259–263
 eddy viscosity
 electromagnetics 3:50–52
 multiscale method 3:40–41, 3:43–44
 Second Moment Closure 3:515
 shallow water equations 3:241
 Smagorinsky 3:40–41, 3:43–44
 subgrid-scale modeling 3:285–287
 turbulence closure 3:304–306, 3:308, 3:310–311, 3:315,
 3:318–322
 variational multiscale method 3:47–49
 Eddy-Break-Up model 3:521
see also large eddy simulations
 edge...
 contributions 1:95
 crossover 3:70
 length quality 1:509
 modes method 1:122, 1:123
 saturation 1:508
 singularities 3:158
 tracked interface locator technique (ETILT) 3:562–564
 EDICT *see* enhanced-discretization interface-capturing technique
 EDMRT *see* enhanced-discretization mesh refinement technique
 EDSTT 3:561, 3:566
 EDSUM *see* enhanced-discretization successive up-date methods
 EEME *see* explicit elliptic momentum equation
 effective...
 bulk modulus 2:419
 indices 3:203, 3:205
 material response properties 2:407–427
 media theory 2:514, 2:515–529
 shear modulus 2:418, 2:419
 stiffness matrices 1:332
 stress principle 2:561
 tensors 2:449–451, 2:454, 2:455
 efficiency
 adaptive computation 1:691, 3:197–198
 airbrake 3:432–433
 explicit residual error estimates 1:88–89
 finite element control 1:86
 vanishing 3:408–410
 velocity evaluation 3:141–145
 EFG *see* element free Galerkin
 eigen-modes 1:200–201, 1:206–207, 1:224–227
 eigen-pair expansions 1:206–207

- eigenfrequencies 1:224–229, 2:766, 2:767
 eigenstrain concepts 2:412
 eigenvalues
 composite laminates 2:456–458
 iteration 1:571–574
 Kötter shell theory 1:217–218
 linear algebraic solvers 1:551–575
 Maxwell equations 1:726–727
 multigrid methods 1:593
 Orthogonal Iteration Method 1:569
 Power Method 1:567–568, 1:569
 QR method 1:567, 1:568–571
 Rayleigh quotients 1:567–568
 Signorini-type interfaces 1:398–399
 tensor fields 1:537
 eigenvectors 1:537, 1:552–553
 EINST 3:566–567
 elastic...
 arterial walls 2:606–607
 axisymmetric soil layers 2:581–582
 bars 1:128–130
 body deformations 2:7–9
 damage 2:344–345, 2:457
 deformation maps 2:244–250, 2:275–276
 domain 2:273
 dumbbells 3:492
 energy 1:200, 1:209–210
 fields 2:727
 ideally plastic materials 2:295–296
 linear damage 2:344–345
 loading/unloading 2:235–236
 plastic crystals 2:280–283
 plastic tangent moduli 2:279–280
 predictors 2:467, 2:473
 property upscaling 2:515–516, 2:521–522
 shakedown
 definition 2:297
 discrete models 2:301–304
 extremum principles 2:299–301
 kinematics 2:294–302, 2:305, 2:318–319
 restrained blocks 2:322–323
 tangent moduli tensor 2:14
 trial states 2:245–246
 viscoelastic correspondence principle 2:751, 2:753–754
 viscous stress splitting (EVS8) 3:483–487, 3:489–490
 elasticity
 based damaged models 2:336–338
 boundary integral equations 2:735–740
 buckling 2:142–144, 2:145–149, 2:150–164
 buried arch structures 2:559–560
 collocation 2:723–727
 coupled BEM/FEM 1:376–377, 1:403–408
 fast solution techniques 2:729–731
 finite element methods 2:5–7, 2:24–46
 geomechanics 2:552
 ligament mechanics 2:627–629
 linear theory 2:7–40
 meshfree methods 1:301–302
 mixed finite element methods 1:241–246, 1:268–269
 nonlinear theory 2:7–16
 shakedown 2:294–304, 2:305, 2:318–319, 2:322–323
 shape sensitivity 2:740–743
 structural mechanics 2:40–42
 symmetric Galerkin BEM 2:727–729
 tensors 2:414–416, 2:449–451, 2:455–456, 2:522–524
 elastodynamics
 dual reciprocity BEM 2:758–760
 nonsingular BEM 2:762–768
 space-time boundary integrals 1:708
 structural dynamics 2:170–171
 elastomeric bead compression 2:480–481
 elastoplasticity
 contact mechanics 2:205, 2:219
 cylindrical billet upsetting 2:479
 damaging solids 2:481–483
 deformations 1:227–264, 2:267, 2:270–271
 augmented Lagrangian 2:260–263
 closest-point-projection 2:251–255, 2:258–259, 2:262–263
 infinitesimal models 2:232–239
 integration 2:244–250
 return-mapping 2:244–250, 2:275–276
 dynamic contact 2:220
 geomechanics 2:552–553, 2:555–556, 2:566–567
p-finite element method 1:133–135
 symmetric Galerkin BEM 2:727–729, 2:738–740, 2:744
 tangents 2:235–237
 elastoviscoplasticity 2:483–484
 electric...
 conductors 1:726
 fields 1:709, 1:724
 fluxes 1:724
 potentials 1:8
 electrodynamics 1:708–709
 electroglvanized steel sheets 2:474–476
 electromagnetism 3:50–52, 3:451
 electron microscopy 2:518–520
 electrostatics 1:724
 ELED *see* essentially local extremum diminishing
 element...
 by element preconditioners 1:566
 extinction methods 2:396–397
 free Galerkin (EFG)
 arbitrary crack growth 2:384–387
 finite element methods 1:504, 1:506
 meshfree methods 1:280, 1:286–289, 1:291, 1:304, 1:306
 partial differential equations 1:291
 Green's function 3:21–25, 3:26, 3:29–30
 nodal interpolation operators 1:80–82
 Pelet numbers 3:24
 shape quality 1:509
 technology 2:476–481
 elter condition 1:252–254, 1:255, 1:257, 1:258–259
 ellipsoid tensor fields 1:537–538
 elliptic...
 boundary integral equations 1:704, 1:705, 1:711
 boundary values
 coupled BEM/FEM 1:375–408
 error estimates 1:95–97, 1:109–111
 finite element methods 1:95–97, 1:107–111
 finite volume schemes 1:465–466
 Galerkin boundary elements methods 1:347, 1:358–366
h-finite elements 1:73
 numerical integration 1:107–109
 Ritz–Galerkin methods 1:74–77
 condition 1:249–254
 conservation laws 1:175, 1:181–187
 continuum damage models 2:342–344, 2:355–362
 damage mechanics 2:342–344
 difference equations 3:341
 hyperbolic problems 1:3

elliptic... (continued)
 mesh generation 1:421
 parabolic problems 1:3
 partial differential equations 1:12–18
 projections 1:692–693
 regularity 1:76–77
 relaxation 3:317, 3:318, 3:321
 shells 1:212–214
 space-time boundary integrals 1:705, 1:711
 variational problems 1:623–627
 embedded anchor bolts 2:360–361
 embedded discontinuities 2:351–355
 embedded multiscale hierarchies 2:423–424
 energetic ordering 2:415–416
 energy
 arbitrary Lagrangian–Eulerian equations 1:419–420, 1:426
 balance 1:331–333
 bending 1:209, 1:215
 boundary 2:744
 buckling 2:140–141, 2:143, 2:147, 2:162
 cascade 3:285
 channel flows 3:212
 complementary Dirichlet energy principle 2:24
 complementary energy functions 2:251
 conservation 1:30, 1:419–420, 1:426, 2:11–13
 consistency 2:445–449
 decaying time integration 2:186–187
 dissipation 2:230–231, 3:308
 elasticity balance equations 2:11–12, 2:13
 elastoplastic deformations 2:230–231, 2:234
 error indicators 3:174–175
 forced homogeneous turbulence 3:238–239
 fracture 2:350–351
 free 2:271–272, 2:273, 2:456, 2:465
 functions 2:201
 harvesting 3:72
 Hill's condition 2:409, 2:411–412
 homogeneous turbulence 3:239–240
 kinetic 2:162
 membrane 1:209, 1:215
 Navier–Stokes equations 3:210
 norms
 error estimates 1:85, 1:86–97, 2:25–26
 forming processes modelling 2:486–487
 generalized 2:486–487
 shallow water equations 3:256–257, 3:258–259
 preserving time integration 2:185–187
 Principle of Minimum Complementary Potential Energy 2:422
 reacting flow equations 3:505, 3:514
 release rates 2:337, 2:456–458, 2:531–532, 2:734
 shallow water equations 3:241–242, 3:256–257, 3:258–259
 shear 1:209
 space-periodic flows 3:211
 spaces 1:350, 1:353, 1:119
 transfer 3:51
 turbulence closure 3:308
 viscoplastic deformations 2:230–231
see also kinetic; potential; strain
 engineering artifact modeling 1:475–494
 Engquist–Other schemes 3:36, 3:39
 enhanced...
 assumed strain 2:93
 continuum models 2:355–362
 discontinuous Galerkin methods 3:33–94
 discretization

fluid dynamics 3:560–561, 3:566–567
 interface-capturing technique (EDICT) 3:560–561
 mesh refinement technique (EDMRT) 3:560–561
 successive up-date method (EDSUM) 3:568–570
 solution techniques 3:566–567
 strain 1:245–246, 1:267–268, 2:93
 ENO *see* essentially nonoscillatory
 entrophy 3:210, 3:211, 3:239–240
 entropy
 conservation laws 1:441–442, 1:447–448
 discontinuous Galerkin 3:97, 3:98–100
 elasticity balance equations 2:12–13
 entropy-entropy flux pairs 1:441
 finite volumes 1:441–442, 1:447–448
 flux pairs 1:441
 inequality 1:447–448, 2:12–13
 envelope surfaces 1:487–489
 environmental acoustics 3:5–7
 equations
 coarse scales 3:16–17
 of equilibrium 2:11, 2:611
 evolution 1:158, 1:169–175, 2:14
 fluid dynamics 3:547–548
 of motion 1:316, 2:161–162
 of state 3:417–418, 3:439, 3:506–507
see also individual entries
 equilibrium estimators 1:91
 equilibrium
 approximations 3:314–315
 boundary layers 3:291–292
 critical states 2:139, 2:144–149, 2:156–159, 2:161–164
 elastic shakedown 2:305
 equations 2:11, 2:611
 operators 2:294–295
 stability 1:672
 states 2:139–149, 2:156–159, 2:161–164
 equivalence concepts 3:509–510
 equivalent plastic strain rate 2:506
 equivalent single-layer (ESL) models 2:436–437
 error...
 absolute 3:96, 3:97
 accumulation 1:676
 adaptive control 1:677–680
 aliasing 3:282, 1:149
 of approximation 1:188
 boundary elements methods 1:339–371
 bounds 3:143–144
 constitutive equations 2:642
 contact mechanics 2:219
 control 1:86–87, 3:170–175
 controlled discretization 2:5–7, 2:37–39, 2:40–46
 elastodynamics 2:766, 2:767
 element nodal interpolation operators 1:80–82
 estimation
 advective-diffusive equations 3:35–36
 anisotropic 1:81–82, 1:63–66
 compressed foam-rubber tubes 2:53–54
 conservation laws 1:445–450, 1:467–468
 energy norms 1:85, 1:86–97, 2:25–26
 explicit 1:87–91, 1:96, 2:28–29
 finite elasticity 2:52
 finite elements 1:86–87, 1:89–93, 1:96, 1:97–98
 frequency 1:448
 fundamental 1:448
 goal functionals 1:85, 1:97–98

goal-orientated 1:97–98, 2:26–27, 2:33–37, 2:48–50
h-finite element spaces 1:68–70
h-symmetric BEM/FEM 1:382–389
 implicit 1:89–91, 1:96
 incompressible viscous flows 3:170–173
 interpolation 1:80–82, 1:512, 1:61–68
 linearized elasticity 2:5–7, 2:24–40, 2:42–44
 Maxwell equations 1:735
 mesh adaptation 1:511–516
 moving boundaries meshing 1:519
 neutron transport equation 3:94, 3:96
 scale separation 3:230–231, 3:234
 symmetric advective-diffusive equations 3:39, 3:40
 failure analysis 2:487–488
 Helmholtz equations 2:696, 2:699–700, 2:701–702
 indicators
 energy 3:174–175
 forming processes modeling 2:485–488
 incompressible viscous flows 3:174–175
 Schur complement 1:384–387
 inf-condition 1:274–276
 interpolation 1:80–82, 1:512, 1:61–68
 linearized elasticity 2:26–27
 nested iteration 1:592
 numerical 3:281–283
 partitioning 2:424–425
 quadrature 1:179–180, 1:603
 reduction property 1:91–93, 1:103
 representation 2:25–27, 2:47–48, 3:190–191
 roundoff 1:656–657
 tolerance 2:158
 truncation 1:52–53, 1:41, 1:48, 3:139–140
 waveform advection 3:62–63
see also a posteriori...; a priori...; averaging...; discretization...; residual...
 Isahelby formalism 2:412
 ESL *see* equivalent single-layer
 essential boundary conditions 1:293–298, 1:299, 1:305–306
 essentially local extremum diminishing (ELED) schemes 3:333
 essentially nonoscillatory (ENO) schemes 1:453–455, 3:354–356
 ETILT *see* edge-tracked interface locator technique
 ETW *see* European transonic wind-tunnel
 EUGENIE 3:413–414, 3:439, 3:449
 Euler
 aerodynamic computations 3:429, 3:432, 3:433
 angles 2:119–120
 codes 3:413–416, 3:439, 3:449
 continuum mechanics 1:413–414
 equations
 acoustic fields 2:697
 compressible 3:101, 3:103–105, 3:383–386
 computational aerodynamics 3:327, 3:329
 gas dynamics 3:101, 3:103–104, 3:105, 3:113–117
 shock capturing 3:348–359
 Euler–Lagrange equations 1:267
 General Galerkin discretization 3:188–189
 homogeneous function theorem 2:238
 implicit time-stepping schemes 3:369
 motion 1:415–416
 operators 1:481–482, 1:483
 parameters 2:119–120
 predictors 1:666
 strain tensors 1:136
 solvers 3:411–412
 visualization algorithms 1:534

European crew transfer vehicles 3:434–435
 European transonic wind-tunnel (ETW) 3:431–433
 evolution equations 1:158, 1:169–175, 2:14
 EVSS *see* elastic viscous stress splitting
 exact...
 blending 1:125–126
 formulations 2:76–79, 2:83
 Maxwell equations 1:727–732
 potential flow equation 3:342–345
 variation equations 3:16–17
 excavations 2:566–567
 existence conditions 1:650–651, 2:640
 exothermic reactions 1:668–669
 experimental techniques, concrete mechanics 2:516, 2:517–520
 explicit...
 algebraic Reynolds stress model (BARSIM) 3:421–423, 3:430
 elliptic momentum equation (BEM) 3:487, 3:489–490
 error estimates 1:87–91, 1:96, 2:28–29
 finite difference method 2:546, 2:547–548
 integration 2:220
 solution methods 2:401–495, 2:546, 2:547–548
 strong stability preserving time integration 1:465
 time integration 1:51, 2:183–184, 2:188–189, 3:252–254, 3:256–259
 time stepping 1:683–686, 1:687, 3:366–368
 total variational diminishing schemes 1:451
 explosions 1:426, 1:427
 exponential...
 convergence 3:110
 decay 1:19–20
 mapping 2:248–250, 2:467–468
 return-mapping 2:248–250
 extended finite element method (XFEM) 2:399–400
 extrusion
 arterial walls 2:609–613, 2:614
 bounded 1:637–638
 exterior...
 acoustics 2:702–703
 boundary values 1:734–735
 Dirichlet boundary conditions 3:9
 displacement 1:345
 generalized Green's formula 1:354, 1:357
 traction 1:345–346
 external loads 1:202, 205
 extinction methods 1:538–541, 2:396–397
 extraction 1:538–539
 extreme eigenvalues 1:398–399
 extremum diminishing schemes 1:445, 3:348, 3:349–351, 3:363
 extremum principles 2:299–301
 extrusion 2:500–502
 F-bar-patch methods 2:478–479, 2:500–502
 F-differentiable mapping 1:650, 1:656
 FA-18 aircraft 3:468–469
 F7X business jets 3:427, 3:431–433, 3:442–443
 FA *see* fully adjusted
 face modes 1:123
 facet flip 1:509–510
 factors of safety 2:305, 2:551, 2:568
 failure
 composite laminates 2:431–458
 error indicators 2:487–488
 loads 2:138

- failure (*continued*)
 modeling 2:335–370
 cohesive-zone models 2:337, 2:349–355, 2:363–365
 discrete 2:335, 2:362–370
- Falcon...
 50 aircraft 3:411, 3:427, 3:433–434
 900 aircraft 3:453
 business jets 3:427, 3:431–433, 3:442–443
 F-16 Block-40 fighter aircraft 3:474–477
 F7X business jets 3:427, 3:431–433, 3:442–443
 falling cubes in water 3:604–607
 falling spheres 3:486–487, 3:571–573
 far-field...
 boundary conditions 3:362
 expansions 1:601, 1:604–605
 partitioning 1:601–602
 scales 3:8–9
- Faraday's law 1:723–724
- fast...
 multiplication 1:600–601, 1:602–604, 1:607, 1:611
 multipole method (FMM) 2:750–751, 3:141–144
 solvers 1:589–594, 2:729–731
- fatigue 2:511, 2:580, 2:206
- Favre averaged values 3:522–523
- FDM *see* finite difference methods
- FEAM *see* finite element alternating method
- FEM *see* finite element method, finite element methods
- FETI *see* finite element tearing and interconnecting
- fiber-reinforced composite laminates 2:431–458
 classical laminate plate theory 2:433–434
 continuum damage mechanics 2:440–441, 2:452–458
 damage 2:431–458
 failures 2:431–458
 layerwise laminate theory 2:435–437
 literature reviews 2:437–440
 macroscale 2:440–451
 microscale 2:440–451
 multiscale modeling 2:432–433, 2:440–451
 partition-of-unity 2:367–368
 plates 1:207, 1:211, 2:433–435
 progressive damage modeling 2:451–458
 representative unit cells 2:443–446
 representative volume element 2:432, 2:441–443, 2:445–451
 shear deformations 2:434–435
- FIC *see* finite calculus
- field...
 elimination 2:580
 equations 2:200, 2:696–697, 3:462
 vortices 1:536
- FieldView application 1:546
- fill-ins 1:565
- filter functions 3:42
- filtered Navier–Stokes equation 3:42–43
- filtered velocity 3:41
- filtering
 large eddy simulations 3:272–274
 length scales 3:278–279
 material responses 2:412
 meshes 3:277–279
 operators 3:272–274
 spectral methods 1:150
- fine grids 3:233
- fine scales
 diffusivity 3:25–26
 Green's function 3:18–20
 variational multiscale method 3:11–27
- fine staggered meshes 3:223–225
- Finger tensors 3:492
- finite calculus (FIC) 3:581, 3:583–587
- finite deformations
 composite laminates 2:441–442, 2:447–448
 contact mechanics 2:196
 elastoplasticity 2:239–244, 2:248–250
 finite difference methods (FDM) 1:3, 1:7–52, 3:360–361
 aerodynamics 3:409–411
 backward 1:36–37, 1:50, 1:51
 convection–diffusion equations 1:43–44
 Crank–Nicolson 1:51
 diffusion equation 1:18–28
 discretization 1:3, 3:360–361
 Du Fort–Frankel–Saul'ev 1:51
 elliptic partial differential equations 1:12–18
 five-point approximations 1:50
 forward approximations 1:50, 1:51
 fourth-order hyperbolic equations 1:34–36
 geomechanics 2:545–546, 2:547–548, 2:562–563
 heat conduction 1:18–28
 hyperbolic equations 1:28–36
 meshfree 1:290
 Navier–Stokes equations 3:213–214, 3:220, 3:228
 parabolic equations 1:18–28
 shallow water equations 3:245–246, 3:248–249, 3:250, 3:251
 soil consolidation 2:562–563
 two-point boundary problems 1:9–12
 variable sensitivity 2:667
 wave equation 1:28–34
- finite dimensions 1:246–257, 2:17, 2:18–20
- finite elasticity
 contact mechanics 2:201
 error estimates 2:46–54
 finite element methods 2:5–7, 2:16–24, 2:46–54
 ligament mechanics 2:627–629
 nonlinear boundary values 2:16–24
- finite elements
 advective-diffusive equations 3:34, 3:39–40
 alternating method (FEAM) 2:381
 anisotropic 2:20–21, 1:57, 1:63–66
 arbitrary Lagrangian–Eulerian methods 1:414, 1:421, 1:431
 arterial walls 2:616–618
 boundary elements symmetric coupling 1:377–389, 1:403–405
 composite laminates 2:431–458
 concept 1:77–79
 conforming 1:73, 1:77–85
 discontinuous 3:91–123
 discontinuous Galerkin methods 3:91–123
 discrete element coupling 1:325–326, 1:334–335
 discretization
 aerodynamics 3:360–361
 buckling 2:140, 2:163
 concrete mechanics 2:534–535
 diffusion 3:170–171
 error estimation 2:5–7, 2:40–46
 error-controlled 2:5–7, 2:37–39, 2:40–46
 forming processes modeling 2:470–471
 Helmholtz equation 2:699
 hyperelasticity 2:16–24
 Navier–Stokes equation 3:160–166, 3:173–174
 penetration detection 2:215–216
 ship hydrodynamics 3:587–589
 spatial 3:160–166
 stochastic 2:663–667

- thin domains 1:220–221
 thin-walled structures 2:108–113
 elastic shakedown 2:303–304
 embedded discontinuities 2:351–355
 equations 1:586–589
 error estimates 1:104–105
 forming processes 2:469–471
 geomechanics 2:549–551
 geotechnical engineering 2:568–569
h-version 1:56–58
 heat wall mechanics 2:622–624
 knee joints 2:628–629
 mesh representation 1:546–547
 meshfree methods 1:293, 1:303–306
 methods (FEM) 1:3, 1:4, 1:73–114, 3:3
 adaptive 1:98, 1:103, 1:510–516, 2:386–390, 2:485–491
 blending meshfree methods 1:293, 1:303–306
 blood flow 3:537–538
 boundary element coupling 1:375–409
 elastoplasticity 2:744–745
 symmetric coupling 1:377–389, 1:403–405
 buried arch structures 2:559–561
 coarse scales 3:15–18
 complex geometric configurations 3:346
 curved domains 1:109–111
 Dirichlet-to-Neumann boundary condition 2:707
 discretization 2:5–7, 2:40–46
 error estimation 1:82–98, 2:5–7, 2:24–54
 finite elasticity 2:5–7, 2:16–24
 fluid dynamics 3:545–574
 fracture mechanics 2:378–381
 generalized 2:399–400
 geomechanics 2:545–547, 2:559–561, 2:564–565
 Hellinger–Reissner 2:21–23
 hierarchical matrices 1:610, 1:612–613, 1:615
 Ha–Washizu 2:23–24
 hybrid 2:21–24
 interior estimates 1:111–112
 kinematic nongeometric fracture mechanics 2:399–400
 linearized elasticity 2:5–7, 2:16–40
 local mesh refinement 1:98–104
 MATLAB 1:98, 1:113–114
 Maxwell equations 1:6, 1:723–736
 mesh refinement 1:98–104
 model adaptivity 2:5–7, 2:24–40, 2:44–46
 panel clustering 1:597, 1:607, 1:610, 1:612–613, 1:615
 partial differential equations 1:292–294, 1:295
 pollution effects 1:111–112
 Ritz–Galerkin methods 1:74–77
 soil consolidation 2:564–565
 superconvergence 1:112–113
 thin domains 1:219–229
 three-field 2:23
 time 2:176, 2:178, 2:184, 2:186
 unbounded domains 2:702
see also coupled...; *hp*-version...; mixed...; *p*-version; stochastic...;
- Navier–Stokes equations 3:583–584
 nodal interpolation error estimates 1:61–63
 non-Newtonian flow 3:490–491
 nonconforming 1:105–107, 1:588–589
 nonuniform meshes 3:565–566
 numerical integration 1:107–109
 parabolic differential equations 1:676, 1:689–690
 plates 1:207, 2:59, 2:79–83
- shakedown analysis, tubes 2:325–326
 shape functions 2:365–369
 shell theory 2:59, 2:79–83
 space-time 2:171–173, 3:27–28
 spaces
 construction 1:73, 1:77–82
 discrete estimates 1:82
 discretization 3:160–161
 element nodal interpolation 1:80–82
 interpolation error estimates 1:80–82
 multigrid methods 1:588
 partitions 1:79–80
 properties 1:77–82
 triangulations 1:79–80
 tearing and interconnecting (FETI) method 1:618–619, 1:641–643, 2:696, 2:712
 thin-walled structures 2:59, 2:104–128
 time-discontinuous space-time equations 2:171–173
 turbulence 3:228
 vascular solid mechanics 2:616–618
 viscoelastic fluid flows 3:484–485, 3:490–491
see also coupled...; *h*-version
- finite hyperelasticity 2:16–24, 2:46–54
 finite layer solutions 2:563–564
 finite plasticity 2:267–287
 finite point methods (FPD) 1:291
 finite spheres 1:280, 1:291–292
 finite strains
 contact discretizations 2:210–214
 crystal plasticity 2:288
 domain boundary element method 2:740
 elasticity 1:431, 2:242, 2:464–467
 exponential return-mapping 2:248–250
 forming processes modeling 2:477–478, 2:481–484
 low-order elements 2:477–478
 viscoplastic deformations 2:242
- finite volume elements (FVE) 1:466
- finite volume (FV) methods 1:4, 1:439–470
 conservation laws 1:439–450, 1:464–470
 discontinuous Galerkin methods 3:97
 discretization 3:360–361
 elliptic boundary values 1:465–466
 Euler code 3:413–414
 flow field compression 1:169–170
 higher-order accuracy 1:450–464
 higher-order time integration 1:464–465
- first...
 law of thermodynamics 2:12
 ply failure (FFF) 2:437–438
 type hexahedral elements 1:728–729
 type tetrahedral elements 1:729–731
- first-order...
 hyperbolic equations 1:28
 monotone schemes 3:419–420
 ordinary differential equations 3:31–32
 reliability method (FORM) 2:678
 saddle points 1:183
 shear deformations 2:86, 2:105–106, 2:434–435
- five-parameter shell models
 degenerating solid elements 2:80
 direct approach 2:77
 shell theory 2:84, 2:87, 2:91–92, 2:94–95
 five-point differencing 1:59, 1:50
 fixed support stencil reconstructions 1:462

- flames
acoustic interactions 3:503–504
flamelet regime 3:517–518
speeds 3:511–512
stretch 3:512–513
flat plate bending 2:64–102
flexible boundaries 1:323–324
flexural cylindrical shells 1:224, 1:226–228, 1:229
flexural displacements 1:202
flexural shells 1:215
flight tests 3:475–476, 3:429, 3:430
flc 3:346
flc22 3:327, 3:328
floating wood 3:603, 3:604
flow
air 2:584–586
bifurcations 1:672–673
cavity flows 3:209, 3:223–225
channel 3:209–210, 3:212–213, 3:221–223, 3:227–228
creep 2:526
cross 1:425–426, 3:487–488
deforming porous medium 2:584–586
field compression 1:169–175
free shear 3:426
Geostrophic 3:253–254
Hermès space shuttle 3:433–434
inviscid flows 3:131–137
 J_2 2:228, 2:239, 2:246–248, 1:134, 1:138
mean 3:302–303, 3:304, 3:318, 3:319, 3:321
non-Newtonian 3:481–496
nondivergent 3:309
past cylinders 3:122–123, 3:201–205, 3:487–488, 3:494–496
periodic 3:219–220, 3:221, 3:235–240, 3:377–379
plasma 3:81–85
plastic 2:274–275
potential flow methods 3:334–348, 3:580
Prandtl–Batchelor 3:133
Quasi–Geostrophic 3:253–254
reacting flow control 3:499–523
separation 3:320
shear 3:294
shock-free transonic 3:346
simulations 3:570–574
soil consolidation 2:564–565
solid boundaries 3:145–149
spatially-periodic 3:210–212, 3:555–557, 3:573
steady 3:358–359, 3:481, 3:482–488
subsonic 3:101, 3:102–104, 3:134–337, 3:425–426
supersonic 3:101, 3:102–104, 3:425–426
theory 2:228, 2:239, 2:246–247
three-dimensional 3:78–81, 3:134–137
two-dimensional 3:77–78, 3:130–134
unbounded inviscid flows 3:131–137
viscous 3:130, 3:137–139, 3:330–331, 3:585–587
wall bounded 3:426
see also blood...; compressible...; computational...; fluid...;
incompressible...; transonic...; viscoelasticity...
fluctuation fields 2:281
fluid...
dynamics
arbitrary Lagrangian–Eulerian 1:421, 1:422–426
meshfree methods 1:280
moving boundaries 3:3, 3:545–574
turbulence closure 3:301–322
flow 3:1–3
aerodynamics 3:330–334, 3:335
complex multidimensional domains 3:559–365
discretization 3:359–365
equations 3:587–588
viscoelasticity 3:3, 3:481–496
fluid-mesh/structure time integrators 3:471–473
flux-driven test cases 2:583–584
object interactions
falling spheres 3:571–573
spatially periodic flows 3:555–557, 3:573
subcomputation technique (FOIST) 3:558–559
pressure loadings 2:694–685, 2:689
rigid-body interactions 1:424, 1:425–426
ship interactions 3:589–590
structure interactions 2:683–692
arbitrary Lagrangian–Eulerian 1:421, 1:423–424, 1:426, 1:427
hydroelastic-clothing 2:683–684, 2:689–692
nonlinear aeroelasticity 3:460–477
structural-acoustics 2:683, 2:684–689
vibrations 2:683
flutter
aerodynamic shape optimization 3:379
computational flow mechanics 3:448–449, 3:450
dual time-stepping 3:376–377, 3:378, 3:379
nonlinear aeroelasticity 3:459–460, 3:464, 3:469, 3:477
fluxes
corrected transport 3:329
difference splitting 3:356
functions 1:169–170
Godunov finite volume discretizations 1:444–445
splitting 3:329, 3:356
thermal diffusion 1:261–262
vector splitting 1:469–470, 3:329
flying conditions 3:379
FMD see fast multipole method
FOIST 3:558–559
Fokker–Planck equation 3:492
foldpoints 1:669–672
focusing tests 2:557–558, 2:559
force...
coefficients 3:373
displacement law 1:316
driven tests 2:583
forced vibrations 2:759
forebody control 3:436
forging 2:496–502
FORM see first-order reliability method
forming processes modeling 2:461–507
adaptive finite element methods 2:485–491
bulk forming operations 2:496–502
contact-friction 2:471–476
continuum constitutive modeling 2:462–469
element technology 2:476–481
explicit solutions 2:491–495
forging 2:496–502
friction modeling 2:471–476
implicit finite element solutions 2:469–471
inelastic constitutive models 2:481–484
meshes 2:488–489
metal cutting operations 2:502–506
strip stretching 2:491
thermomechanical coupling 2:484–485
thin sheets 2:495–496
transfer operators 2:489–491
Fortin's trick 1:270–272
forward finite differencing 1:50
forward-backward McCormack finite differencing 1:51
forward-time, central space finite differencing 1:51
foundation and error analysis 1:539–571
four-point bending 2:39–40
Fourier...
Bessel series 2:563–564
coefficients 3:243–244
expansions 1:34–35
Galerkin approximation 3:214–215
Legendre polynomials 3:215–217
representations 3:377–379
series 1:716, 2:563–564
spectral methods 1:141–142
transforms
boundary integral equations 1:704, 1:711–712, 1:713–714
pressure 3:6–7
soil consolidation 2:563–564
space-time boundary integrals 1:711–712
time-harmonic waves 2:698
fourth-order equations 1:34–36, 1:85
FPD see finite point methods
FFF see first ply failure
fractal boundaries 1:654
fractional time derivatives 2:752–753
fractional-step methods 3:218–219, 3:588–589
fracture 2:375–402
boundary element method 2:732–735
concrete mechanics 2:532–533
discontinuous deformations 1:330–331
discrete crack models 2:362, 2:363
discrete element methods 1:329–331
energy 2:350–351
see also computational fracture mechanics
fragmentation 1:329–331
Fredholm derivative 1:182
Fredholm differential mapping 1:650, 1:656
Fredholm integral equations 1:180, 1:594
Fredholm theorem 1:10, 1:352, 1:354
free...
charge density 1:724
energy
continuum damage mechanics 2:456
crystal plasticity 2:271–272, 2:273
elasticity balance equations 2:13
potentials 2:465
fixed rods 2:757
shear flow 3:426
stream sensitivity 3:309
surfaces
arbitrary Lagrangian–Eulerian dynamics 1:423, 1:424
boundary conditions 3:580, 3:581–582, 3:586–587, 3:589
flow 3:574
liquids 2:689
wave equation 3:586–587, 3:589
vibration analysis 2:759
Frelat's method 2:319–320
frequency
domains 1:704, 1:713–714
eigenfrequencies 1:224–229, 2:766, 2:767
errors 2:179
multifrequency methods 2:710–711
zero-frequency mode 2:686, 2:691
friction
arbitrary Lagrangian–Eulerian 1:431, 1:433
cohesive-frictional soils 2:550–551
contact
constraints 1:315, 1:328
friction 2:471–476
mechanics 2:201–202, 2:204–206, 2:219–220
Coulomb 1:315, 1:328, 1:433, 2:205–206
dry 2:205
dynamic contact 2:220
forming processes 2:471–476
frictionless contact 2:200–201
skin 3:307, 3:315, 3:321–322
slip corrector 2:473
tangential slip 2:206
Tresca 1:433
velocity 3:307
Probenius hierarchical matrices 1:614
Proude number 3:580
frozen active sets 2:276–278
fully...
adjusted (FA) states 2:554
discrete finite volume methods 1:443–444
populated matrices 1:597–615
saturated soils 2:589–592
functional subgrid-scale modeling 3:284, 3:285–287, 3:290
fundamental aerodynamic studies 3:436–437
fundamental error estimates 1:448
fundamental solutions
boundary integral equations 1:707–709, 1:716–717
displacement integrals 2:720–721
elastodynamics 2:766
space-time boundary integrals 1:707–709
fusion 3:115, 3:117
FV see finite volume
FVE see finite volume elements
G-NT see Galerkin with Numerical Integration
Galerkin
boundary element methods 1:347, 1:358–366
Aubin–Nitsche lemma 1:364–365
Céa's lemma 1:358–362
convergence optimality 1:362–364
elliptic boundary values 1:347, 1:358–366
ell-positiveness 1:365–366
panel clustering 1:600
stability 1:365–366
boundary values 1:24–25
Bubnov–Galerkin weak form 1:292–293
discretization 1:650, 1:606–607, 1:612, 3:170
general 2:700–701, 3:184–185, 3:187–189, 3:191–198
equations 1:358–362
finite element methods 2:699, 3:583
finite volume methods 3:413–414
Fourier–Galerkin approximation 3:214–215
least-squares (GLS)
advective-diffusive equations 3:34–35, 3:36, 3:37, 3:39, 3:40
Helmholtz equations 2:700–701
Hermès space shuttle 3:433–434
Navier–Stokes code 3:417
stabilized methods 3:22–23
methods
adaptive wavelets 1:177–180
boundary integral equations 1:346
conforming 1:177

- Galerkin (*continued*)
 coupled BEM/FEM 1:376, 1:379–382, 1:385
 Dirichlet-to-Neumann boundary condition 2:707
 Euler code 3:413–414
 Helmholtz equations 2:698–699, 2:700–701
 Navier–Stokes equations 3:214–217
 space-time boundary integrals 1:709–712
 subgrid-scale models 3:30–32
 time discretization 3:170
 time-discontinuous 2:173, 2:176–177, 2:184
see also discontinuous...;
 with Numerical Integration (G-NI) 1:143–144, 1:147, 1:149
 orthogonality 2:25, 2:47
 projections 1:207, 1:210
 Ritz–Galerkin 1:73, 1:74–77, 1:556, 1:557–558
 Taylor–Galerkin 3:583
 wavelet schemes 1:178–180
 weak forms 1:291–293
 weights 1:361
see also element free...; Petrov–Galerkin; Runge–Kutta...;
 symmetric Galerkin...
 Gallian invariance 3:275
 Gauss finite difference method 1:521–52
 Gårding inequalities 1:351–352, 1:369–370, 1:379
 gas dynamic motion 1:426–427, 3:101, 3:103–105, 3:113–117
 Gauss law 1:724
 Gauss–Lobatto integration 1:143–144
 Gauss–Lobatto points 1:121
 Gauss–Seidel implicit time-stepping 3:370–372
 Gauss–Seidel smoothers 3:178–179
 Gaussian...
 distribution 2:559–660
 elimination 1:563–564
 filters 3:273–274
 window functions 1:281
 General Galerkin discretization 2:700–701, 3:184–185, 3:187–189,
 3:191–198
 Generalized Linear Model 3:313
 generalized...
 energy norms 2:486–487
 explicit total variational diminishing schemes 1:451
 finite difference 1:290
 finite element method (FEM) 2:399–400
 Galerkin method 2:700–701, 3:184–185, 3:187–189, 3:191–198
 Green's formula 1:354–355, 1:357
 Hermite polynomials 2:674–676
 implicit total variational diminishing schemes 1:451–452
 minimal residuals (GMRES) 1:558–559, 2:702, 3:565–566, 3:570
 real Schur form 1:570
 representation formula 1:355–358
 self-consistent method 2:412
 serial staggered (GSS) procedure 3:473
 slope limiters 3:107–108
 geomaterials 2:551–558, 2:559
 geomechanics 2:543–569
 consolidation 2:543, 2:561–567
 constitutive models 2:551–558, 2:559
 deterministic analysis 2:544–545
 limit analysis 2:549–551
 numerical analysis methods 2:545–549
 soil-structure 2:558–561
 stochastic techniques 2:567–569
 geometric...
 coconservation laws 1:424–425, 3:466–468
 consistent integration 2:274–275
 continuum constitutive forming process modeling 2:463–464
 deformations 1:519
 error estimation 1:519
 exact formulations 2:76–79, 2:83
 extraction 1:538–539
 modeling 1:5
 aerodynamic shape optimization 3:438–439
 attributes 1:490–492
 blood flow 3:538–540
 boundary representation 1:477–485, 1:487–489
 complex shapes 1:475–494
 constructive solid geometry 1:483–485
 engineering artifacts 1:475–494
 medial modeling 1:485–490
 surface patches 1:477–481
 system architecture 1:475–476
 turbulent flame combustion 3:519, 3:520
 voxel representation 1:2–3, 1:10–11, 1:476–477
 nonlinearity 2:310–311, 1:135–136
 representation 2:378–394
 shell theory 2:76–79, 2:83, 2:93–95
 three-dimensional shells 2:83–102
 visualization 1:526
 geophysics 3:251–259
 Geostrophic flows 3:253–254
 Germano identity 3:274–275, 3:288–289
 GFEM *see* generalized finite element method
 glass-fiber-reinforced polypropylene 2:361–362
 global...
 contact mechanics 2:197, 2:198, 2:216–220
 continuity projection-based interpolation 1:733
 interpolation error estimates 1:68–70
 a posteriori error estimates 2:27–33
 searches 2:196, 2:214–216
 GLS *see* Galerkin least-squares
 glyphing 1:530, 1:533, 1:538, 1:540–541
 GMRES *see* generalized minimal residuals
 goal functionals 1:85, 1:97–98
 goal-orientated error estimates 1:97–98, 2:26–27, 2:33–37, 2:48–50
 Godunov...
 finite volume discretizations 1:442–445
 fluxes 3:99
 like convection phase stress-updates 1:430–431
 gradients
 aerodynamic shape optimization 3:440–442
 based optimization methods 2:644
 conjugate iterations 1:642–643, 1:147
 convective projections 3:586
 damage mechanics 2:355–362
 definition 3:308
 deformation 2:8–9, 2:66, 2:463, 2:477–478
 discontinuities 1:302–303
 displacement 1:244–245, 2:343–344
 enhanced damage models 2:356–359, 2:369
 iterations 1:185, 1:642–643, 1:147
 mesh deformations 3:441–442
 modified deformations 2:477–478
 objective functions 2:644
 operator 1:727–732
 plasticity-damage models 2:359–360
 recovery estimates 1:93–95, 1:96–97
 reduced formulation 3:386–387
 smooth particle hydrodynamics 1:284–285
 Sobolev 3:388, 3:389
 graft-artery bypass junctions 3:536, 3:537
 Gram matrices 1:287
 granular media 1:316, 1:323–324
 graphics 1:527–528
 gravitational potentials 1:8
 gravity 2:689–692, 3:247–248, 3:251
 green water 3:581
 Green–Gauss linear reconstructions 1:461–462
 Green–Lagrangian strain tensors 2:66–67, 2:73–74, 2:91, 2:441–442
 Green–Naghdi stress rate 2:494–495
 Green's formula 1:354–355, 1:357
 Green's function
 coarse scales 3:14–15
 convection–diffusion equations 1:44–45
 difference schemes 1:44–45
 Dirichlet boundary conditions 3:9–10
 element 3:21–25, 3:26, 3:29–30
 fine scales 3:18–20
 space-time formulations 3:29–30
 variational multiscale method 3:18
 vortex methods 3:130, 3:141
 grids
 computational 1:547–548
 filters 3:272–274
 fine 3:233
 hierarchies 1:580
 particle-grid method 3:144–145
 sub-grid stress 3:32–33
 two-grid iterations 1:581–584
see also coarse...; multigrid methods; subgrid-scale models
 ground effects 3:432, 3:433
 ground transportation systems (GTS) 3:148, 3:149
 growth 2:615–616, 2:622
see also crack...
 GSS *see* generalized serial staggered
 GTS *see* ground transportation systems
 GUI-based frameworks 1:544
 Gunton model 2:341
 plasticity 2:553
 shakedown 2:307–310
 hardware operation levels 1:476
 harmonic Ritz values 1:553
 Harten–Lax–van Leer flux 1:470
 Harten's explicit total variational diminishing schemes 1:451
 Haskin–Shrikkan bounds 2:412–413, 2:418, 2:420
 hat-functions 1:395
 haunched continuous plates 2:44–46
 Hausdorff topologies 1:487
 head 1:520–521, 2:565
 headlamp panels 2:496, 2:497
 heart valves 2:624–625
 heart wall mechanics 2:618–625
 heat...
 conduction 1:18–28, 1:622
 equations
 adaptive computation 1:676, 1:689–692, 1:696–699, 1:714–719
 error estimates 1:690–691
 logistics reaction-diffusion 1:697–698, 1:699
 space-time boundary integrals 1:706, 1:710–712, 1:713
 space-time Galerkin finite elements 1:676, 1:689–690
 strong stability estimates 1:691–692
 time-stepping 1:696–697, 1:714–719
 flux vectors 2:12
 potential operators 1:718–719
 release 3:499
 transfer
 Second Moment Closure, turbulence 3:315
 thermo-hydrodynamics 2:592–599
 turbulence closure 3:320, 3:321
 vortex methods 3:149
 Heaviside-type traction 2:762, 2:763
 heavy liquids 2:688–689
 hedgehogs 1:533
 helicity 3:130–131
 helicopters 3:104, 3:105, 3:376–379, 3:571, 3:572
 Hellinger–Reissner functional
 dual-mixed finite element method 2:21–23
 elasticity 1:242–243, 1:244, 1:256, 1:268–269
 thermal diffusion 1:239–240, 1:257–262
 Helmholtz equation
 accelerated multifrequency methods 2:710–711
 acoustics 2:697–702
 boundary integral equations 1:705, 1:716
 elasticity 1:242–243, 1:244, 1:256, 1:268–269
 Dirichlet-to-Neumann formulation 3:8–9
 discretization 2:696, 2:698–702
 element Green's function 3:26
 space-time boundary integrals 1:705
 time-harmonic waves 2:697–698
 wavenumbers 2:699–701, 3:26
 Helmholtz free energy 2:13
 Helmholtz operator 3:18
 hemodynamic conditions 3:537–538
 Hencky elasticity 1:376–377, 1:403–405
 Hencky strain energy function 2:466–467
 Hencky strain tensors 2:9
 Hencky–von Mises stress-strain relation 1:377
 Hencky–von Mises type materials 1:376–377, 1:403–405
 Hermite spaces shuttle 3:412, 3:433–434
 Hermite elements 1:77–78
 Hermitian matrices 1:552
 Hermitian polynomials 2:1213–214, 2:671–676
 Hertz–Signorini–Moreau conditions 2:201
 heterogeneity 2:281–283, 2:286–287, 2:432–433

hexahedral...
 elements 1:728–729, 2:38, 2:214
 shape functions 1:123–124
 to tetrahedral transformations 3:65
 hidden variables, Maxwell equations 1:727
 hierarchical...
 error estimators 1:93, 2:31–32, 2:35–37, 2:52
 interpretations 2:421–425
 matrices 1:597, 1:607–615
 models
 adaptive 2:425
 elasticity 2:40–42
 mesh layout 1:223–224
 plates 1:201–202, 1:207–211, 1:223–224
 shells 1:201–202, 1:218–219, 2:106–107
 structural mechanics 2:40–42
 p -refinement 3:18–20
 polynomial expansions 3:65–67
 shape functions 1:733–734, 1:120 124
 two-level decompositions 1:385–386
 High...
 angle of attack 3:460, 3:429–431, 3:436
 cycle fatigue 2:311
 level operation levels 1:476
 lift effects 3:432, 3:433
 resolution switched schemes 3:251
 speed machining 2:502–504, 2:505
 speed trains 3:5–7, 3:570–571
 temperature hypersonic flows 3:433–434
 High Irradiance RESponse (HIRES) 1:685, 1:686
 higher fundamental solutions 1:716–717
 higher-dimensional manifolds 1:666–667
 higher-dimensional parameter spaces 1:671–672
 higher-order...
 accurate finite volume methods 1:450–464
 continuum models 2:355–362
 elements 2:112–113
 elliptic partial differential equations 1:13–15
 explicit time-stepping schemes 3:367–368
 gradient models 2:355–362
 hierarchical models 1:210–211
 meshes 1:530
 models 1:210–211, 2:106–107, 2:355–362
 nonoscillatory shock capturing 3:353–354
 polynomial approximations 2:700
 predictors 2:155
 Stokes elements 3:164
 time integration 1:464–465
 Hilbert spaces 1:74, 1:349–350, 1:363
 Hill's condition 2:409, 2:411–412
 HIRES *see* High Irradiance RESponse
 historical overviews
 aerodynamics 3:525–529, 3:408–413
 computational mechanics 1:1–2
 domain decomposition 1:619–621
 mesh generation 1:498–499
 plate theories 2:60
 thin-walled structures 2:59–60, 2:71–72
 hodograph transformation 3:381
 Hölder inequality 1:67
 Hölder spaces 1:342–343, 1:356
 homogeneous...
 elastic-plastic deformations 2:270–271
 functions, Euler theorem of 2:238
 material properties 2:85

plastic deformation 2:269–271
 reactors 3:513–514
 turbulence
 isotropic 3:209–240
 Navier–Stokes equations 3:228–240
 spatial/time behaviors 3:227
 two-level decomposition 3:228–232
 homogenization
 axial stress–axial strain curves 2:342
 composite laminates 2:432–433, 2:439–440
 concrete mechanics 2:515–516, 2:522–524
 crystal plasticity 2:269–271, 2:281–282, 2:286–287
 elastic-plastic crystals 2:281–282, 2:286–287
 micromechanics 2:407–427
 multiscale modeling 2:407–427
 homotopies 1:489–490, 1:661–662, 1:665
 Hood–Taylor elements 1:263–264, 1:266
 Hooke materials 2:14
 Hookean dumbbells 3:492–495
 Hooke's laws 1:207, 1:208, 2:15–16, 2:582
 hoop strains 2:87–89
 Hopf bifurcations 1:672–673
 horizontal motion, ships 3:602, 3:603
 hot jets 3:445–446
 h -p convergence method 1:126–131
 h -adaptivity 1:104, 1:723–736, 3:91, 3:101–102
 h -clouds 1:291–292
 h -finite element methods 3:1
 coarsening strategies 1:104
 coupled boundary element 1:389–394
 polynomial expansions on unstructured grids 3:63–67
 spectral elements 3:61–88
 Hsieh–Clough–Tocher macro elements 1:78
 Hu–Washizu finite element method 2:23–24
 Hu–Washizu functional 1:240, 1:243–244, 1:245–246, 1:256
 hull waters 3:581
 hulls 3:581, 3:589–590, 3:594–599
 hybrid methods
 constitutive equations 2:644–645
 convection–diffusion equations 1:47–50
 displacement boundary elements 2:763–766
 Eulerian/Lagrangian solvers 3:149
 finite element 2:21–24
 incomplete LU factorization 1:563–566
 Lainer tunnel 2:527–529
 Schwarz theory 1:626, 1:627
 stress finite element 2:24
 tunnel linings 2:514, 2:516–529
 hydration 2:526
 hydrocodes 1:422
 hydrodynamics, ships 3:579–607
 hydroelastic sloshing 2:683–684, 2:689–692
 hydrostatic boundaries 1:323–324
 hygral property upscaling 2:516
 hyperbolic...
 advective-diffusive equations 3:39
 boundary integral equations 1:705, 1:706, 1:711, 1:712, 1:713
 conservation laws 1:466–470, 1:169–175
 difference equations 3:341
 finite difference equations 1:28–36
 introductory survey 1:3
 space-time boundary integrals 1:706, 1:711, 1:712, 1:713
 time-stepping methods 1:715
 hyperbolicoid flares 3:427–428

hyperelasticity
 constitutive relation 2:13–14
 contact mechanics 2:196–197, 2:200–201, 2:202
 elastic constitutive model 2:465–467
 geomechanics 2:552
 hyperelastoplasticity 1:429–430
 hypersingular...
 boundary integral operators 1:349
 kernel integral equations 1:594
 operators 1:576, 1:176, 1:177
 hypersonic flows 3:433–434
 hyperviscosity 1:538, 1:539
 hyperviscosity 3:241
 hypocoelasticity 2:552
 hypocoelasticity 1:428–430
 hypoplasticity 2:553
 ICF code 3:115, 3:117
 idealization, geomechanics 2:544–545
 identification experiments 2:517–520
 identifying material parameters 2:637–654
 ignition times 3:511, 3:513
 ill-posedness 1:365–366, 2:640–641
 image analysis 2:517–520
 image-order volume rendering 1:542
 impacts 1:280, 2:379
 impedance boundary conditions 1:726
 imperfect bonding 2:439
 imperfection sensitivity 2:149–150, 2:349
 implicit...
 constitutive integration 2:736–740
 error estimators 1:89–91, 1:96
 finite element solution 2:469–471
 functions 1:538–539
 residual error estimators 1:387–389, 2:29–31, 2:34–35, 2:50–52
 solution methods 2:546
 temporal discretizations 3:488–490
 thickness integration 2:83
 time integration 2:182–183, 2:188–189, 3:252
 time-stepping schemes 3:368–372
 total variational diminishing schemes 1:451–452
 imploding simulation 3:115, 3:117
 impressed surface currents 1:726
 in-plane material models 2:535–536
 in-plane strains 2:92
 incomplete Cholesky decompositions 1:562, 1:565
 incomplete LU factorization 1:562–567
 incompressible...
 elasticity 1:244–245, 1:254–256
 flows 3:2
 computability/adaptivity 3:186–187
 convection–diffusion 3:121–122
 energy-harvesting 3:72
 inertialless 3:482–488
 isothermal viscoelastic fluid 3:482
 mean flow 3:304
 ship hydrodynamics 3:581–583
 spatial discretization 3:69
 stabilization parameters 3:549–550
 Taylor vortex 3:71–72
 time integration 3:67–69
 viscosity 3:2, 3:155–179
 error control 3:170–175
 mathematical models 3:156–160

space discretization 3:160–166
 time discretization 3:166–170
 vortex methods 3:131–137
 wake 3:72–74
see also turbulence...
 hydroelastic sloshing 2:683–684, 2:689–692
 meshfree methods 1:298–300
 Navier–Stokes equations
 computability/adaptivity 3:186–187
 finite element methods 3:160–166, 3:548–549
 multiscale method 3:40–55
 space-time formulation 3:44–45
 spatial discretization 3:160–166
 time dependent domains 3:69 70
 incremental...
 boundary values 2:469–470
 collapse 2:304–307, 2:323, 2:325
 constitutive law 2:469
 crack growth simulation 2:377
 stability 2:280
 variational formulation 2:280
 indefinite finite elements 1:104–105
 indices
 laminar flow around cylinders 3:203, 3:205
 orientation 1:607–608
 reliability 2:677
 Sobolev index 1:366–371
 indirect methods 2:157
 industry
 aerodynamics 3:407–454
 p -finite element method 1:134–135, 1:136–137
 inelastic...
 constitutive equations 2:638–640
 cracked bodies 2:311
 distortions 2:267
 materials 2:481–484, 2:638–640
 inertial...
 confinement fusion simulation 3:115, 3:117
 equations 2:173–174
 inertialless incompressible flows 3:482–488
 manifolds 3:211–212
 range of scales 3:211–212
 waves 3:251
 inelastic...
 Additive Schwarz Method 1:623–625
 multiplicative Schwarz methods 1:625–626
 Newton's methods 1:660–661
 subdomain solvers 1:636–638
 inextensional displacement p 1:215
 inf-condition
 mixed finite element methods 1:248–252, 1:254, 1:257–259, 1:266–276
 proof techniques 1:269–276
 saddle-point stability 1:248–249, 1:251–252, 1:254, 1:257
 Stokes equations 1:266–267
 thermal diffusion 1:258–259
 viscoelastic fluid flows 3:484
see also Babuška–Brezzi condition
 infinite dimensional problems 1:159, 1:185–186
 infinite elements 2:708–710, 2:711–712
 infinitesimal deformations 2:232–239, 2:245–248, 2:442–443, 2:448–449
 infinitesimal elastoplasticity 2:232–239, 2:245–248
 inflation 2:609–613, 2:614
 inflow boundary conditions 3:291, 3:292–293

information visualization 1:526
infrared (IR) signatures 3:451–452
Inglebert's method 2:319–320
initial value problems (IVPs) 1:676–702, 2:268–269
inlet designs 3:428–429, 3:433, 3:434
integrals
 Cauchy 1:343, 1:344–345, 1:718–719
 conservation equations 1:419–420
 conservation law 1:442
 continuous boundary operators 1:351
 differentiation 2:741
 longitudinal scale 3:211
 multigrid methods 1:593, 1:594–595
 nonlinear multigrid iterations 1:593
 panel clustering 1:599
 strain interior points 2:735–736
 stress interior points 2:735–736
 time derivatives 1:418–419
 transforms 2:564
 volume 1:418–419
 see also boundary integral equations
integration
 cells 1:293
 contact-friction model 2:473
 degenerating solid elements 2:83
 elastoplasticity 2:244–250, 2:736–740
 explicit 1:465, 2:220
 Gauss–Lobatto 1:143–144
 plastic flow 2:274–275
 semi-implicit time 3:245, 3:246, 3:252, 3:254–256
 semi-Lagrangian time 3:69
 stress 2:467–468, 2:493–495
 through-the-thickness 2:54
 viscoelastic fluid flow 3:490–491, 3:492–493
 viscoplastic deformations 2:244–250
 visualization algorithms 1:534–535
 see also numerical ...; time ...
integrity tensors 2:454–456
Intelligent Light FieldView application 1:546
inter penalty (IP) method 3:119–119
interacting body characterization 1:317–321
interfaces
 boundary conditions 2:697
 capturing 3:560–562
 computational systems 1:546–548
 elements 2:363–365
 fluid dynamics 3:3
 Signorini-type 1:376–377, 1:396–403
 tracking 3:561–562
 visualization systems 1:546–548
interior ...
 displacement 1:344
 estimates 1:111–112
 points 2:735–736
 traction 1:344
interlaminar failure 2:439
intermediate mortar surfaces 2:210
internal ...
 energy 2:11–12
 length scales 2:361–362
 modes 1:122–123
 variable dependency 2:310
 variable mapping 2:489–490
 work 2:446–449

interpolation
 arbitrary Lagrangian–Eulerian 1:421–422
 errors 1:180–82, 1:512, 1:61–68
 estimates 1:30–82, 3:35, 1:61–68
 h -finite element spaces 1:55–70
 interpolated sub-time stepping technique 3:561
 meshes 1:421–422, 1:516
 mismatches 1:547
 nodal 1:80–82, 1:58–59, 1:61–66
 operators 1:30 82, 1:54, 1:58 60
 projections 1:732–734
 Stokes equations 1:264–266
 viscoelastic fluid flows 3:484
 see also approximation
intersections 1:302, 2:114–117, 1:133
intrinsic material functions 2:526–527
invariance
 affine 1:657–658
 Galilean 3:275
 Lax–Wendroff scheme 3:415
 Navier–Stokes equations 3:275
 positive streamwise 3:415, 3:420–421
 reflection 3:275
 rotational 1:733–734, 3:275
 time shift 3:275
inverse ...
 constitutive equations 2:640–641
 hierarchical matrices 1:611–612, 1:615
 Laplace transforms 1:707
 power method 2:159
 stiffness matrices 1:611–612
inviscid ...
 equations 3:331–332
 flows 3:131–137
 flux terms 3:82–84
 pitching cycles 3:380
IP *see* inter penalty
IR *see* infrared
irregular boundary domains 3:157–158
irregular data sets 1:529
isocontours 1:532
isolas 2:160
isoparametric ...
 contact discretizations 2:207
 elements 1:731–732
 finite element method 1:109
 interpolations 2:209, 2:724
 mapping 1:58
 surface elements 2:212–213
isothermal solutions 2:589–592
isotropic ...
 bound meshes 1:513
 compressible materials 2:14–16
 compression 2:555–556
 damage mechanics 2:336–338
 elasticity 2:205–206, 2:726
 finite elements 1:61–63
 finite hyperelasticity 2:16–24
 finite strain 2:242, 2:248–250
 function representation 2:244
 linear elasticity 2:205–206
 material properties 2:85
 Maxwell equations 1:723–736
 plasticity 2:243–244, 2:249
 simplices 1:57

turbulence
 closure 3:312, 3:313
 direct numerical simulations 3:294
 homogeneity 3:209–240
 large eddy simulations 3:295–296
 Navier–Stokes equations 3:209–240
 yield criterion 2:235
Isotropization of Production (IP) 3:313
iteration
 adaptive wavelets 1:185–187
 arbitrary Lagrangian–Eulerian mechanics 1:431
 contact mechanics 2:197–198, 2:216
 coupled BEM/FEM 1:376
 defect correction 3:176
 direct linear algebraic solvers 1:555–556
 Dirichlet-to-Neumann boundary condition 2:707–708
 eigenvalues 1:571–574
 fluid dynamics 3:565–566
 Iterated Standard Partitioned Procedure 2:583
 Maxwell equations 1:735
 multigrid methods 1:580–581, 1:590–592
 nonlinear parabolic equations 1:26
 nonlinear systems 1:650–661
 substructuring 1:633–635
IVPs *see* initial value problems

J_2 -flow theory 2:228, 2:239, 2:246–248, 1:134, 1:138
Jackson estimate 1:164, 1:167
Jacobi ...
 Davidson method 1:574
 implicit time-stepping 3:370
 iteration 1:581–583
 matrices 1:593, 3:331–332
 orthogonal polynomials 1:143–146
 preconditioner 1:390–391
 Jacobians 2:200
Jameson–Schmidt–Turkel (JST) scheme 3:529, 3:531, 3:533, 3:537, 3:563
Jummann stress rate 2:493–494
jets 3:398–399, 3:427, 3:431–433, 3:436–437, 3:442–448
Johnson–Mercier elements 1:268, 1:269
Jordan matrix forms 1:553
JST *see* Jameson–Schmidt–Turkel
jumps
 boundary elements 1:351
 buckling 2:160, 2:161–164
 discontinuous Galerkin methods 3:92–96, 3:103
 operators 2:172
 residuals 1:57–88
 space-time boundary integrals 1:706

 k - ϵ model 3:303–307, 3:320, 3:321
 k - l model 3:306–307
 k - w model 3:308–310, 3:320, 3:321
Karhunen–Loeve expansion 2:665–667, 2:671–676
kernels
 ellipticity 1:252–254
 expansions 1:606–607
 explicit constructions 1:717–718
 far-field expansions 1:605
 meshfree arbitrary shape approaches 2:383–384
kinematics
 admissible displacements 1:209–210
 arbitrary Lagrangian–Eulerian method 1:416–417
 arterial walls 2:610–611

classical laminate plate theory 2:433–434
constitutive equations 2:650
constraint 2:339–340
elasticity 2:7–10, 2:294–302, 2:305, 2:318–319
enhancement 2:453
equations 2:91–93
finite elements 2:303–304
forming process modeling 2:463–464
hardening 2:308–309, 2:553
layer discretization 2:112–113
nongeometric representations 2:394, 2:399–400
shakedown 2:294–302, 2:305, 2:318–319
shear deformations 2:434–435
shells 1:219
space enrichments 1:209–210
kinergy energy 2:162
kinetic energy
 aeroacoustics 3:446–447
 decay inequality 3:45, 3:48–49
 elasticity balance equations 2:11–12
 homogeneous turbulence 3:237–240
 shallow water equations 3:241–242, 3:254, 3:255–259
 spectral amplitude 3:43–44, 3:47
 turbulence closure 3:303
kinetics
 classical laminate plate theory 2:433–434
 reacting flow combustion 3:507
 response 2:161
Kirchhoff displacements 1:203–204, 1:229
Kirchhoff stress
 contact mechanics 2:200
 copper single crystals 2:283–284
 Green–Naghdi stress rate 2:494–495
 Jaumann rate 2:493–494
 tensors 2:10, 2:14, 2:230–231
Kirchhoff–Love models
 classical laminate plate theory 2:433–434
 discretization 2:109–110
 plates 1:204, 1:208–209, 1:229, 2:433–435
 shear deformations 2:434–435
 shells 1:211–218, 2:103–105
 Kiware ParaView application 1:546
 knee joints 2:628–629
 known-solution methods 2:381–382, 2:383
 Koiter, Warner T.
 buckling 2:145–147, 2:150
 clamped elliptic shells 1:214
 plates 1:201, 1:217–218
 shakedown theorem 2:297–298
 shells 1:201, 1:217–218, 2:103–105
 see also Kirchhoff–Love
Kolmogorov ...
 energy spectrum 3:43–44, 3:47
 law 3:211–212
 length scales 3:280–281
 Koren limiter 1:457
 Korn airfoils 3:346
 Korn's inequalities 1:75
 Kuznetsov entropies 1:441
 Krylov projection methods 1:556–557
 Krylov subspaces
 accelerated multifrequency methods 2:711
 Dirichlet-to-Neumann boundary condition 2:708
 eigenvalues 1:571–574
 linear algebraic solvers 1:552, 1:556–557, 1:560, 1:561
 Kuhn–Tucker complementary conditions 2:223

Kohn–Tucker–Karush conditions 2:201, 2:218, 2:219
 KVLCC2 hull model 3:595, 3:596, 3:597–599

L-shaped domains 1:131
 L-stability 2:178–179
 L^1 -contraction principle 1:441, 1:447, 1:451
 L^2 -condition number 1:365–366
 L_2 projections 1:692–693, 1:142
 L_2 -norms 2:27–28, 3:261–262
 laboratory tests 2:638–639
 Ladyženskaya–Babuška–Brezzi (LBB) condition
 adaptive wavelets 1:159, 1:181, 1:183
 meshfree methods 1:299–300
 Sobolev index 1:368–370
 viscoelastic fluid flows 3:484
 Lagrangian
 buckling stability 2:142–143
 continuum mechanics 1:413–414
 convection–diffusion equations 3:149
 description of motion 1:415–416
 finite element spaces 1:77
 fluid flow 3:581, 3:591–592, 3:594–607
 multipliers
 contact discretizations 2:206–207, 2:208
 contact mechanics 2:202–203, 2:216–217
 discrete elements 1:322, 1:329
 matrix form 2:208
 meshfree methods 1:293, 1:294–296
 model interpolation error estimates 1:65
 particle distribution 3:136, 3:140–141
 phase stress-update 1:429–430
 Laisner tunnel 2:525–529
 LAM *see* Limited Area Model
 Lamé constants 1:204, 2:67, 2:99
 laminar...
 flames 3:500–501, 3:512–514
 flow around cylinders 3:201–205
 level deformation 2:432
 laminates
 level deformation 2:432
 theories 2:433–437
see also fiber-reinforced composite...
 Lanczos method 1:557, 1:571–574
 Lanczos–Tau method 1:143
 Laplace domains 2:754, 2:755–756
 Laplace transforms 1:704, 1:707, 1:711–712, 1:713–714
 Laplacian smoothing 1:421
 Laplacian vertex placement 1:541
 large data visualization methods 1:542–543
 large eddy simulations (LES)
 adaptive 3:184–185, 3:191–197
 aerodynamics 3:423–426
 applications 3:265–296
 boundary conditions 3:291–293, 3:296
 computability 3:184–185, 3:191–197
 direct numerical simulations 3:184–185, 3:191–198
 eddy viscosity 3:47–49
 filtering operator 3:272–274
 multiscale method 3:40, 3:41–43
 numerical error 3:282–283
 resolution requirements 3:281
 shallow water equations 3:241
 time advancement schemes 3:283
 turbulence 3:2, 3:269, 3:270–274, 3:279–285, 3:291–293,
 3:295–296

turbulent flames 3:515–517, 3:520
see also direct numerical simulations
 large strains 2:566, 2:649–654, 3:309–310
 Large Time Increment (LATIN) 2:582–584
 large-scale decompositions 3:219–228
 large-scale separations 3:248–251
 lateral boundary conditions 1:207
 lateral wave contact 3:604–605
 LATIN 2:582–584
 lattices 2:269–271, 2:391–392, 2:339–340
 Launder, Reece and Rodi (LRR) turbulence closure 3:313
 laws
 Ampère's 1:723–724
 Arrhenius 3:517
 constitutive 2:310, 2:204–206, 2:621, 2:624
 Darcy's 2:582
 Fourier's 1:723–724
 first law of thermodynamics 2:12
 force displacement 1:316
 Gauss 1:724
 Hooke's 1:207, 1:208, 2:15–16, 2:582
 Kolmogorov 3:211–212
 law-of-the-wall 3:307
 law-of-the-wall 3:307
 material 2:66, 2:74, 2:84, 2:97, 2:98–102
 second law of thermodynamics 2:12–13
see also conservation
 Lax–Friedrichs finite difference 1:51
 Lax–Friedrichs flux 1:470, 3:55, 3:59
 Lax–Milgram lemma 1:348, 1:352, 1:353–354
 Lax–Wendroff...
 explicit time-stepping 3:366–367
 finite difference 1:51
 flux 3:414–415
 positive strainwise invariance 3:415
 theorem 1:447, 3:98
 layer-wise models 2:106–107, 2:435–437
 LBB *see* Ladyženskaya–Babuška–Brezzi
 LCD *see* limit cycle oscillations
 LDG *see* local discontinuous Galerkin
 leap-frog schemes 1:51, 2:183–184, 2:188–189, 3:252–254,
 3:256–259
 least squares
 adaptive wavelets 1:185
 arbitrary crack growth 2:384–386
 boundary integral equations 1:346, 1:347
 centered moving 1:288–289
 complex geometry 3:346–348
 continuous moving 1:285–286
 coupled BEM/FEM 1:376, 1:394–396
 Helmholtz equations 2:701
 incompressible viscous flows 3:176
 linear reconstructions 1:461
 meshfree methods 1:280, 1:285–291
 moving 1:280, 1:285–291, 2:384–386
 nonlinear systems 1:651, 1:657–658
 parameter identification 2:643, 2:645–646
 stabilizations 3:187
see also Galerkin...
 LED *see* local extremum diminishing
 LEFM *see* linear elastic fracture mechanics
 left-preconditioning 1:561
 Legendre polynomials 1:393, 2:704, 1:143–144
 Lemaitre's damage model 2:488, 2:503

length scales
 cohesive-zone models 2:350
 concrete mechanics 2:520–521
 determination 2:361–362
 Navier–Stokes equations 3:210–213, 3:218
 parameters 3:592–593
 shells 1:216, 1:223–224
 Leonard tensors 3:274
 Leonard's decomposition 3:274
 LES *see* large eddy simulations
 Lesaint–Raviart method 3:483
 level of loading 2:514, 2:517–520, 2:527, 2:533–534, 2:537–538
 level-of-detail (LOD) 1:542
 Liapounov criterion 2:142–143
 Lie derivative 2:243
 lift...
 coefficients 3:84–85, 3:146–147
 effects 3:432, 3:433
 forces 3:203–204
 over-drag 3:451
 lifting operators 1:251–252
 lifting schemes 1:165
 ligaments 2:625–629
 Lightail turbulence tensors 3:6
 lighting models 1:528
 limit
 analysis 2:549–551
 cycle oscillations (LCO) 3:460
 factors 2:299–301
 model 1:208–209, 1:211–218
 points 2:145–147, 2:149, 2:150, 2:156–160
 state boundary 2:568
 state functions 2:676–677, 2:678, 2:679
 Limited Area Model (LAM) 3:242–246
 line relaxation methods 3:340–341
 line-tracked interface up-date technique (LTIUT) 3:564–565
 linear...
 acoustics 2:695–714
 advection equations 1:150
 algebraic solvers 1:5, 1:551–575
 direct methods 1:553–560
 LU-factorization 1:554–555, 1:562–567
 preconditioning 1:555–556, 1:560–562
 combination subgrid-scale modeling 3:287–288
 complementary problems 2:219
 convection–diffusion-reactions 3:185, 3:190
 elastic...
 axisymmetric soil layers 2:581–582
 bars 1:128–130
 continua 2:669–671, 2:673–676, 2:677–678
 fracture mechanics (LEFM) 2:575, 2:581, 2:384–385, 2:394–399
 elasticity 2:7–40
 elastodynamics 2:764–766
 elliptic boundary values 1:74–77
 finite elements 1:288
 hyperbolic equations 3:91, 3:92–96
 Maxwell equations 1:723–736
 momentum 2:11, 2:67
 multigrids 3:177–179
 multistep (LMS) 2:175–176, 2:178
 one-dimensional equations 3:532
 operators 1:168–169, 1:191
 prebuckling state 2:144–145, 2:147–149
 reconstruction (LP) 2:316–317, 2:549–551
 reconstructions 1:457–462

regression continuous discretization 2:664–665
 shell theory 2:102
 stability 3:489–490
 structural dynamics 2:181–184
 symmetric hyperbolic equations 3:91, 3:94–95
 vibratory response 2:683–692
 linearization
 buckling 2:141, 2:144–145
 contact-friction 2:473–476
 duality 1:693–694
 forming processes modeling 2:471
 incompressible viscous flows 3:176–179
 local convergence theory 1:657–658
 Navier–Stokes code 3:417
 primary state 2:141, 2:144–145
 Reynolds averaged turbulence 3:421
 shakedown 2:312–313
 linearized elasticity
 adaptive mesh refinement 2:57–39
 collocation 2:734
 error estimates 2:27–37
 error estimation 2:5–7, 2:24–40
 finite element methods 2:5–7, 2:16–40
 hybrid finite element methods 2:21–24
 material responses 2:408, 2:427
 mixed finite element methods 2:21–24
 model adaptivity 2:5–7, 2:24–40
 nonlinear boundary values 2:16–24
 Schwarz domain decomposition 1:622–623
 symmetric Galerkin BEM 2:729
 Lims' lemma, domain decomposition 1:629
 Lipschitz...
 domains 1:167
 estimate in time 1:447
 surfaces 1:710–711
 liparification 2:554
 liquid filled tubes 3:571–573
 liquid motions 2:692
 literature overviews
 adaptive computation 1:676–678
 composite laminates 2:437–440
 multigrid methods 1:579
 a posteriori error control 1:73, 1:87
 thin-walled structures 2:61
 LMS *see* linear multistep
 load...
 buckling 2:140, 2:152–153
 concrete mechanics 2:533–534, 2:538
 control 2:126
 deformation space curves 2:152–153
 displacement curves 2:545, 2:537–538
 distributions 3:380
 domains 2:296–297, 2:298–299, 2:300–304
 factors 2:305–306, 3:475–476
 geomechanics 2:561
 intensity 2:140
 multistep soil dynamics 2:592, 2:593, 2:594
 parameter 2:152–153
 plates 1:202–205
 transverse 2:438
 vectors 1:528
 loading–unloading conditions
 concrete mechanics 2:514, 2:517–520, 2:527, 2:533–534, 2:537–538
 damage mechanics 2:338
 microgeometrical manufacturing 2:416
 plastic 2:235–236, 2:340–341

- loading-unloading conditions (*continued*)
 shakedown 2:293–294, 2:296–297, 2:300–304, 2:321–326
 tensile 2:437–438
 thermal 2:301
 thermomechanical 2:293–294, 2:321–326
- local...
 conservativity 3:92
 contact mechanics 2:197
 convergence theory 1:649–669
 coordinates 3:64–65, 3:66
 curvature estimates 1:515–516
 discontinuous Galerkin (LDG) 3:117–118, 3:120–123
 discrete maximum principle in space 1:445–446
 errors 1:699, 1:701
 extratum diminishing (LED) schemes 1:445, 3:348, 3:349–351, 3:363
 Green's functions 1:44–45
 interpolation error estimates 1:61–68
 mesh refinement 1:98–104
 parameterizations 1:664–666, 1:669–673
 residuals 3:93–94, 3:96, 3:103
 space-time discrete maximum principle 1:445–446
 stiffness matrices 1:113–114
 time derivatives 1:418
- locality
 adaptive wavelets 1:166
 implicit error estimators 1:90–91
 projection-based interpolation 1:733
- locally upwinded spectral technique (LUST) 3:483
- locking
 discretization 2:111
 finite element formulation 2:120–125
 forming processes modeling 2:476
 meshfree methods 1:298–300
 thin domains 1:221–223
- LOD *see* level-of-detail
- log-layer solutions 3:305, 3:306, 3:307
 log-normal random variables 2:660–661
 logarithmic strain measures 2:464–465
 logical scriptable attributes 1:492
 logistics reaction-diffusion 1:697–698, 1:699
 longitudinal integral scale 3:211
- loosely coupled fluid/fluid-mesh/structure time integrators 3:471–473
- Lorenz system 1:680–683
- low...
 level operation levels 1:476
 order elements 2:477–478
 Reynolds number k - ϵ model 3:306
 speed effects 3:432, 3:433
 lower bounds 1:97–98, 2:414
- LP *see* linear programming
- LRR *see* Launder, Renne and Rodi
- LSC stabilization dynamics 3:550
- LSSD stabilization flows 3:173–174
- LTIUT *see* line-tracked interface up-date technique
- LUI-factorization 1:554–555, 1:562–567
- Lacy states 1:520–521
- Liders band propagation 2:346, 2:348–349
- lumped parameter models 3:530–531
- LUST *see* locally upwinded spectral technique
- Lysapunc equation 1:615
- Mach numbers
 aerolasticity 3:460, 3:475–476
 fluid flow 3:334
- shock-capturing 3:359
 subgrid-scale modeling 3:290
- macroelements 1:78, 1:266
 macromechanics 2:440–451, 2:452–458
 macroscale composite laminates 2:440–451
 macroscopic...
 constitutive behavior 2:449–451
 continuum slip theory 2:267–268, 2:269–270
 damage 2:452–458
 free energy 2:271–272
 material models 2:535–537
 material responses 2:407–427
 magnesium 2:648–649
 magnetic fields 1:724
 magnetohydrodynamics (MHD) 3:81 84
 magnetostatic 1:724
 man-made structures 2:558–561
 Mandel–Cryer effect 2:562
 manifolds 1:662–664, 1:666–667, 3:211–212
- mapping
 affine 1:58, 1:62–63
 bump 1:491
 color 1:531–532, 1:533–534
 displacement 1:533–534, 2:247, 2:490–491
 elastic deformation 2:244–250, 2:275–276
 exponential 2:248–250, 2:467–468
 Fréchet-differentiable 1:650, 1:656
 internal variable 2:489–490
 isoparametric 1:58
 p -finite element method 1:124–126
 particle motion 1:417–418
 pitchfork 1:669–670
 return 2:244–250, 2:275–276, 2:219, 2:467–468
 transfinite 2:275–276
- Visualization algorithms 1:531–532, 1:533–534, 1:539
- matching cubes 1:476–477, 1:484–485, 1:532, 1:533
- margins of safety 2:551
- MARK 1:98, 1:99, 1:100
- marking criterion 1:100
- Mars Lander 3:527, 3:528
- mass...
 balance 3:585
 conservation 1:419–420, 1:426, 2:10, 3:581–582, 3:585
 fractions 3:510–511
 matrices 1:571, 2:764, 2:766–767
 transfer 2:594–596, 2:592–599
 master-slave concept 2:210–211, 2:212–213
- material...
 acceleration 1:419
 derivatives 1:418–419, 3:129
 domains 1:414, 1:416, 1:417–418
 elasticity tensors 2:455–456
 frame indifference 2:239–240, 2:241
 functions 2:526–527
 heat flux vector 2:12
 heterogeneity 2:432–433
 instabilities 2:341–349
 laws 2:66, 2:74, 2:84, 2:97, 2:98–102
 layers 2:112–113
 matrices 2:75
 nonlinearities 2:72–76
 parameter identification 2:637–654
 response properties 2:407–427
 setup 2:85–86
 stiffness 2:398–399, 2:553
 tensors 2:66–67, 2:449–451, 2:455–456

- mathematical models
 fluid flow 3:330–334, 3:335
 incompressible viscous flows 3:156–160
 thin-walled structures 2:61–68
 turbulent flows 3:271–279
- mathematical programming methods 2:202, 2:216, 2:219–220
- MATLAB 1:98, 1:113–114
- matrices
 compression 1:175–181
 contact discretizations 2:208–209
 exponential function 1:615
 formulation discretization 1:600
 linear algebraic solvers 1:552
 matrix-matrix multiplication 1:611
 matrix-vector
 addition 1:611
 computation 3:567–568
 multiplication 1:600–601, 1:602–604, 1:607, 1:611
 notation 2:15–16
 subtraction 1:611
 truncation 1:611
- mechanical properties 2:426
 monotonicity 1:16–17
 partitioning 2:580–582
 saddle-point stability 1:248–257
see also stiffness...
- maximum angle condition 1:83
- maximum principles
 discrete 1:10–11, 1:445–446
 elastic shakedown 2:300, 2:302
 finite volume methods 1:445–446, 1:459, 1:463–464
 parabolic equations 1:18–19
- Maxwell compatibility expression 2:343
- Maxwell equations
 de Rham diagrams 1:729, 1:732–734
 discontinuous Galerkin methods 3:94
 exact sequences 1:727–732
 finite element methods 1:6, 1:723–736
 projection-based interpolation 1:732–734
 space-time boundary integrals 1:708–709
 variation formulation 1:725–727
- MCC *see* Modified Cam Clay
- MEDEM *see* modified distinct element method
- MDS *see* multilevel diagonal scaling
- MEAM *see* modified embedded atom method
- mean
 drag coefficient 3:84–85, 3:192, 3:194, 3:195
 flow 3:302–303, 3:304, 3:318, 3:319, 3:321
 shear stresses 3:535, 3:536–537
 value (random variables) 2:659
 value (vectors) 2:646
 velocity 3:301, 3:307
- mechanical...
 linear elastostatics 2:408
 power supply 2:12
 property upscaling 2:516
 stress 2:10
- mechanics
 carotid artery 2:613, 2:614
 composite laminates 2:439
 material responses 2:425–426
 thin-walled structures 2:63–68
see also computational flow; computational fracture; contact;
 continuum; damage; geomechanics; micromechanics
- media theory 2:514, 2:515–529
- medial axis 1:485–490
 medial modeling 1:485–490
 medical imaging data 3:538–540
 Melan–Kotter's theorem 2:297–298
 Melan's theorem 2:297–299
 member (structural) scales 2:530–538
- membranes
 deformation patterns 1:213
 displacements 1:202, 1:206–207
 dominated action 2:68–70
 elements 2:108–109
 energy 1:209, 1:215
 generator 1:204–205
 locking 1:221–222, 2:121, 2:123–124, 2:125
 operators 1:213
 shells 1:215
 theory 2:102–103
- memory limitations/usage 2:198, 3:485–486
- mercury intrusion porosimetry (MIP) 2:517–519
- mesh...
 acceleration 1:419
 adaptation 1:498
 adaptive finite element methods 1:510–516
 arbitrary Lagrangian–Eulerian 1:420, 1:422
 error estimates 1:511–516
 forming processes modeling 2:488–489
 incompressible viscous flows 3:170–175
 unit volume 1:507–510
 aerodynamics 3:359–360, 3:396
 arbitrary crack growth 2:389–390
 bias 2:362–363
 deformation gradients 3:441–442
 density responses 2:417, 2:418
 depending weights 1:57
 discretization 3:362–363
 element construction 1:500
 evolution 2:506
 filtering 3:277–279
 generation 1:5, 1:497–502
 adaptive finite elements 1:510–516
 aerodynamics 3:426–428
 forming processes modeling 2:488–489
- introductory survey 1:1
 layout 1:223–224
 meshfree methods 1:490
 moving boundaries 1:517–521
 node updating 3:590–591
 refinement 1:87, 1:98–104, 2:488
 regularization 1:420–422
 representation 1:546–547
 resolution rules 2:699–700
 sensitivity 2:341, 2:344–349
 smoothing 1:421–422, 1:541
 thin domains 1:220
 topological representation 1:546–547
 updating 1:420–422, 3:553–555, 3:590–591
 velocity 3:70
- meshfree methods 1:3–4, 1:279–306
 approximation 1:280–291
 arbitrary shape approaches 2:383–386
 blending finite element methods 1:303–306
 convection-diffusion equations 1:45–47
 differencing 1:45–47, 1:250, 1:291
 discontinuities 1:300–303
 finite difference method (FDM) 1:290, 1:291

meshfree methods (*continued*)
 higher-order damage models 2:359
 incompressibility 1:298–300
 local Petrov-Galerkin (MLPG) method 1:291
 moving least squares 1:280, 1:285–291
 partial differential equations 1:291–300
 radial basis functions 1:300
 smooth particle hydrodynamics 1:280, 1:281–285, 1:291
 volumetric locking 1:298–300

metals
 aluminum 2:379, 2:417–421, 2:548–549
 cutting operations 2:502–506
 microplane damage models 2:539–340
 stamping 1:518, 1:519
see also steel

meteorology 3:241–242

method of...
 characteristics 1:47–50, 1:439–440
 contours of vorticity curl 3:131, 3:132–133
 finite spheres 1:280
 lines 1:20–21, 1:15, 3:196, 3:167–168
 methodical depths 1:2
 methodical widths 1:2
 metric definition 1:507
 metric structures 1:487
 metric tensors 1:211–212
 MFDOM *see* meshfree finite difference method
 MHD *see* magnetohydrodynamics
 microcracking 2:437–440, 2:526, 2:553
 microgeometrical manufacturing 2:414–416

micromechanics
 composite laminates 2:432–433, 2:437–451, 2:452–458
 concrete mechanics 2:515–519
 homogenization 2:407–427
 multiscale modeling 2:407–427
 normal contact stresses 2:204
 microplane damage models 2:539–340
 micropolar theory 2:76–79, 2:83
 multiscale composite laminates 2:440–451
 microscopic constitutive behavior 2:449–451
 microscopic damage 2:432–433, 2:452–458
 microstructures 2:267–287
 midpoint rule 3:54–55
 midpoint step functions 2:664
 midsurface curvature tensors 1:211–212
 mild steel axisymmetric necking 2:651–652, 2:653–654

military aircraft
 aerelasticity 3:460, 3:461, 3:474–477
 computational flow mechanics 3:428–431, 3:436
 forebody control 3:436
 transonic flow 3:526–327

MINI elements 1:262–263, 1:266, 1:270
 minimal residuals, generalized 1:558–559, 2:702, 3:565–566, 3:570
 minimum angle conditions 1:79–80
 minimum norm residuals 1:556, 1:558
 MINRES 1:389, 1:393

MIP *see* mercury intrusion porosimetry
 Mirage 2000 aircraft 3:429, 3:430

mismatched mesh representation 1:547

MITTCT *see* mixed interface-tracking/interface-capturing technique

mixed derivative elliptic partial differential equation 1:14

mixed element-matrix-based/element-vector-based computation technique (MMVCT) 3:567–568

mixed finite element methods 1:3, 1:237–276
 elastic shakedown 2:303–304
 elasticity 1:241–246, 1:268–269, 2:21–24

inf-condition 1:248–252, 1:254, 1:257–259, 1:266–276
 linearized elasticity 2:21–24
 local discontinuous Galerkin 3:117–118
 saddle-point stability 1:246–257
 stability 1:246–257
 Stokes equations 1:240–241, 1:262–268
 thermal diffusion 1:238–240, 1:257–262
 viscoelastic fluid flows 3:481–496

mixed function subgrid-scale modeling 3:286–287

mixed interface-tracking/interface-capturing technique (MITTCT) 3:561–562

mixed modeling 3:519–520, 3:521

mixed principles 2:300, 2:302, 2:305

mixing enhancement 3:425–426, 3:436–437

mixture fractions 3:508–511

MLPG *see* meshfree local Petrov-Galerkin

MLS *see* moving least squares

MMVCT *see* mixed element-matrix-based/element-vector-based computation technique

modal analysis 1:209–210

mode jumping 2:160, 2:161–164

model...
 adaptivity 2:5–7, 2:24–40, 2:44–46
 attributes 1:490–492
 decisions 2:83–102
 error estimates 2:42–44
 parameters 1:492

models
 continuum failure 2:335–336, 2:341–349, 2:369
 failure 2:335–370
 thin-walled structures 2:59–103
 turbulent flame combustion 3:519–521
 visualization algorithms 1:538–541
see also forming processes; geometric modeling; multiscale methods

modified...
 Cam Clay (MCC) model 2:556–558
 deformation gradients 2:477–478
 distinct element method (DEM) 1:334
 embedded sigma method (MEAM) 2:391–394
 Hellinger-Reissner functional 1:240, 1:243, 1:244
 incomplete LU factorization 1:563
 Lax-Wendroff flux 3:414–415
 nonreflecting Dirichlet-to-Neumann boundary condition 2:705–708
 variational principles 1:293
 wave numbers 3:282

molecular transport 3:505–506

Molenskamp model 2:556

moment...
 closures 3:301, 3:311 318, 3:320
 of linear momentum conservation 2:11
 predictions 3:239

momentum
 arbitrary Lagrangian-Eulerian 1:419–420, 1:426
 contact mechanics 2:200
 equations
 conservation 1:419–420, 1:426
 reacting flow 3:505
 structural dynamics 2:173–174
 turbulence closure 3:302
 viscoelastic fluid flow 3:491–492
 preserving time integration 2:185–186
 ship hydrodynamics 3:551–552, 3:555
 monitoring buckling 2:156
 monolithic methods 2:592–599
 Monot model 2:556

monotonicity
 elliptic partial differential equations 1:15–17
 finite volume methods 1:451, 1:452–453
 Godunov finite volume discretizations 1:444
 monotone integrated large eddy simulations 3:283
 monotone upstream-centered scheme for conservation laws 1:452–453, 1:456–457, 3:329
 turbulence equations 3:419–420

Monte Carlo method 2:646, 2:647, 2:549

Mooney-Rivlin material 2:14

Morawetz's theorem 3:546

Mori-Tanaka scheme 2:524

Morse theory 1:493

mortar method 1:643–644, 2:210, 1:152–154

motion
 arbitrary Lagrangian-Eulerian 1:413–418
 elastic body deformations 2:7–9
 equations 1:516, 2:161–162
 Eulerian 1:415–416
 Lagrangian 1:415–416
 ship hydrodynamics 3:602, 3:603

moving...
 boundaries 1:498, 1:517–521, 3:3, 3:545–574
 domains 3:69–70, 3:76–77
 heat sources 1:696
 interfaces 3:545–574
 least squares (MLS) 1:280, 1:285–291, 2:384–385
 meshes 3:565–566
 reaction fronts 1:698–699, 1:700

MSM *see* multiplicative Schwarz methods

multi-dimensional consolidation 2:562

multi-chemical reactions 3:510–511

multibody systems
 block deformability 1:324–329, 1:334–335
 boundary conditions 1:321–324
 conservation time integration 2:186, 2:187
 contact constraints 1:321–324
 discontinuous deformations 1:326–329
 discrete element methods 1:311–329, 1:331–335
 structural dynamics equations 2:174, 2:186, 2:187
 time integration 1:331–333

multicoloring 1:563–564

multidimensions
 finite volume methods 1:455–464
 hyperbolic conservation laws 3:108–110
 shock capturing 3:363–364

multidirector models 2:106

multidisciplinary field equations 3:462

multidomains 3:412–413

multifield problems 2:575–599

monolithic methods 2:592–599

partition solution procedures 2:577–589

soil dynamics 2:589–592

multifrequency solution methods 2:710–711

multigrid methods 1:5
 additive variants 1:589–590
 boundary element method 1:593–595
 eigenvalues 1:593
 finite element equations 1:586–589
 general remarks 1:577–581
 implementation case 1:578
 incompressible viscous flows 3:176, 3:177–179
 iterations 1:590–581, 1:590–592
 literature 1:579
 Maxwell equations 1:735

nested iteration 1:590–592

subspace iteration 1:590

time-stepping schemes 3:372–375

turbulent flows 3:233–234

two-grid iterations 1:581–584

multilayer formulations 2:112–113

multilevels
 closures 3:288
 component compaction 1:431, 1:432–433
 diagonal scaling (MDS) 1:633
 error estimators 1:91–93
 homogeneous isotropic turbulence 3:209–240
 overlapping domain decomposition 1:632–633
 shallow water equations 3:240–259
 time integration 3:252–259
 turbulence 3:2

multiparameter subgrid-scale modeling 3:289–290

multiphase flow 2:584–586, 2:592–599

multiphysics 2:575–599, 3:412–413

multiple reciprocity 2:726

multiple sample tests 2:418

multiplicative
 decomposition 2:464
 panel clustering 1:600–601, 1:602 604, 1:607, 1:611
 plasticity 2:240–244, 2:248–250
 Schwarz methods (MSM) 1:620–621, 1:625–630

multipoles
 expansions 2:705–707, 3:141–144
 moments 2:730–731
 panel clustering 1:600

multiresolution
 constructions 1:163–164
 geometric modeling 1:480–481
 visualization 1:543

multiscale methods 3:1
 advective-diffusive equations 3:34, 3:37
 composite laminates 2:432–433, 2:440–451
 concrete mechanics 2:513–539
 crystal plasticity 2:267–287
 Dirichlet-to-Neumann formulation 3:8–11
 expansions 1:202–218
 homogenization methods 2:407–427
 incompressible Navier-Stokes equation 3:40–55
 material responses 2:421–425
 micromechanics 2:407–427
 modeling 2:407–427, 2:432–433, 2:440–451, 2:513–539
 space-time formulations 3:27–32
 stabilized methods 3:5–55
 turbulence 3:40–55

multislip crystal plasticity 2:277

multistage explicit time-stepping schemes 3:367–368

multistep time integration 2:175–176

multizone boundary element method 2:732

Murman-Cole difference scheme 3:339–340, 3:341–343

muscular arterial walls 2:606–607

myocardial tissue 2:619–625

NACA *see* National Advisory Committee for Aerodynamics

Naghdi model 1:201, 1:223–224, 2:86

Nanson's formula, elastic bodies 2:8

NASA crew rescue vehicles 3:434–435

National Advisory Committee for Aerodynamics (NACA)
 6 series airfoils 3:326–327
 0012 airfoils 3:78–81, 3:101–103, 3:122–123
 0012 profiles 3:593, 3:594, 3:595

National Advisory Committee for Aerodynamics (NACA) (continued)

0012 wing inverse design 3:394–397
 0012 wing transonic flow 1:519–520
 0015 pitching airfoils 3:77–78
 National Science Foundation 3:473–477
 NATM *see* New Austrian Tunneling method
 natural draught cooling towers 2:533–538
 natural trace spaces 1:363
 Navier–Stokes equations
 aerodynamic code 3:412, 3:416–418, 3:433–434
 aerodynamic shape optimization 3:383–386
 blood flow 3:533, 3:534
 compressible 3:122–123
 computational fluid dynamics 3:183
 dynamic multilevel methods 3:228–240
 finite element discretization 3:173–174
 finite element methods 3:548–549
 homogeneous isotropic turbulence 3:209–240
 LSD stabilization 3:173–174
 multilevel methods 3:209–240
 renormalization theory 3:259–260
 ship hydrodynamics 3:581–584
 shock capturing 3:348–359
 spectral methods 1:146–148
 supercomputing 3:412
 turbulence 3:209–240, 3:271–272, 3:274–277, 3:301
 viscous discretization 3:364–365
 see also incompressible . . . Stokes equations
 near-field partitioning 1:601
 near-field scales 3:8–9
 near-wall modeling 3:305–306, 3:308, 3:315–317, 3:320
 nearly singular integrals 2:725
 necking 2:284
 Nedetz's construction 1:730
 neighbor searches 1:318
 neo-Hooke materials 2:14
 nested finite element spaces 1:587
 nested iterations 1:590–592
 nested solutions 3:172–173
 von Neumann analysis equation 3:345
 Neumann boundary conditions 1:15, 2:67, 3:220–221
 Neumann problems 1:707, 2:29–30, 2:34–35, 2:50–52
 Neumann–Neumann preconditioners 1:618, 1:639–641
 neural networks 2:642–643
 neutral equilibrium state 3:139, 2:144
 neutron transport equation 3:91, 3:92–96
 New Austrian Tunneling method (NATM) 2:525–529
 Newton potential 1:716
 Newton–Raphson algorithm 2:18, 2:471
 Newtonian fluids 3:269–296, 3:304, 3:533
 Newton's methods
 adaptive wavelets 1:186, 1:194
 crystal plasticity 2:278–279
 incompressible viscous flows 3:176
 local convergence theory 1:657, 1:658–659, 1:660–661
 multigrid methods 1:593
 shakedown 2:314–315
 thin-walled structures 2:126–127
 NIP capsule implosion simulation 3:115, 3:117
 nine-point differencing 1:13–14, 1:50
 Nitsche method 1:297–299, 1:364–365, 2:203–204, 2:210
 NLP *see* nonlinear programming
 NMR *see* nuclear magnetic resonance
 no-slip boundaries 3:209, 3:220–221, 3:305, 3:308, 3:316
 nodal interpolation 1:80–82, 1:58–59, 1:61–66

nodal modes 1:122, 1:123
 node splitting 1:626–627, 2:380–381
 node-decoupling 2:380–381
 node-regular refinement 2:37–38
 node-to-segment (NTS) 2:209, 2:211–212, 2:215
 noise 3:444–448
 non smooth . . .
 contact 1:314–316, 1:334
 solutions 1:130
 surfaces 2:213–214
 non-Newtonian fluids 3:2, 3:481–496
 non-self-adjoint problems 1:37–38
 nonassociative constitutive laws 2:310
 nonlinear crack propagation 2:378–379
 nonconforming finite elements 1:105–107, 1:588–589
 nonconservative difference equations 3:341–342
 nondimensional moment predictions 3:239
 nondivergent flows 3:309
 nonflexural modes 1:227, 1:228
 nongeometric representations 2:394–400
 nonisothermally fully saturated consolidation 2:586–589
 nonlinear . . .
 AERO simulation platforms 3:473–477
 boundary values 2:16–24
 computational aeroelasticity 3:439–477
 conservation laws 1:439–450, 1:468–470, 3:91, 3:96–115, 1:150
 constitutive equations 3:310–311
 continua 2:64–70, 2:679–680
 difference equations 3:340–341
 elasticity 1:376–377, 1:403–405, 2:7–16
 elastoplasticity 1:133–135
 equations 1:592–593
 forming processes 2:485–491
 Galerkin method 3:49
 geometric applications 1:135–136
 Hencky–von Mises stress-strain relation 1:377
 hyperbolic problems 1:158
 kinematic hardening 2:308–309
 least-squares 3:176
 loosely coupled fluid/fluid mesh/structure time integrators 3:471–473
 material behavior 2:408, 2:451–452
 multigrid iteration 3:176
 one-dimensional equations 3:532, 3:533
 operators 1:191–193
 parabolic equations 1:23–26
 programming (NLP) 2:317–318, 2:550
 reaction rates 3:499
 softening 2:451–452
 solid mechanics 1:426, 1:428–433
 strong stability preserving time integration 1:465
 structural dynamics 2:184–187
 theory 2:7–16
 three-field formulation 3:461–464
 nonlinearity 1:6
 adaptive wavelets 1:183–184
 buckling 2:141
 geometric 2:310–311, 1:135–136
 local convergence theory 1:649–669
 mathematical modeling 2:61–70
 shells 2:68–70
 soil consolidation 2:566
 nonlocal damage models 2:356–357, 2:358–359
 nonmatching discrete interface compatibility 3:468–471
 nonmatching meshes 2:209–210
 nonmoving meshes dynamics 3:565–566

nonnested finite element spaces 1:588
 nonnormal random fields 2:567
 nonoscillatory schemes 1:453–455, 3:348, 3:349–356
 nonoverlapping domain decomposition 1:91, 1:618–620, 1:633–644, 2:712
 nonpenetration conditions 1:314, 1:328, 2:198–199, 2:201
 nonpremixed flames 3:500–501, 3:502
 nonreflecting Dirichlet-to-Neumann boundary conditions 2:703–708
 nonshakedown 2:299
 nonsingular . . .
 hybrid boundary element method 2:762–768
 integrals 2:725
 symmetric boundary element method 2:762–768
 nonstationary incompressible viscous flows 3:175
 nonstationary Navier–Stokes equations 3:166–170
 nonsliff initial value problems 1:680–683
 nonsymmetric finite elements 1:104–105
 nonuniform large strains 2:649–654
 normal . . .
 contact stresses 2:204
 derivative kernels 1:605
 displacement 2:685–686, 2:690
 distance 2:198–199
 distribution 2:659–660
 frictional contact 2:472
 gap variation 2:199–200
 interfaces 1:424
 stresses 3:310–311
 normalized normal density 2:659–660
 normalized objective functions 2:425
 nnnms
 equivalences 1:166–168
 of the error 3:121–122
 error estimates 1:87, 2:25–26, 2:27
 linear elliptic boundary values 1:74–75
 plates and shells 1:229
 saddle-point stability 1:247–248
 NTS *see* node-to-segment
 nuclear magnetic resonance (NMR) 2:518
 numerical . . .
 codes 3:413–418
 diffusion fluxes 1:467
 discretization 2:416–417
 dispersion numbers 1:33–34
 error 3:281–283
 filters 3:272–274
 fluxes
 conservation laws 3:497, 3:499
 discontinuous Galerkin methods 3:92, 3:95, 3:119
 Euler code 3:414
 functions 1:442–443, 1:469–470
 RKDG 3:109
 second-order elliptic problems 3:117
 implementations
 dynamic multilevel methods 3:231–240
 gradient-enhanced damage models 2:357–358
 renormalized scales 3:260–263
 viscoelastic direct boundary elements 2:757–758
 integration
 collocation 2:724–725
 elastoplastic deformations 2:227–264
 finite elements 1:107–109
 symmetric Galerkin BEM 2:729
 viscoplastic deformations 2:227–264
 visualization algorithms 1:534–535
 layers 2:112–113

methods
 arterial walls 2:611–612
 boundary integral equations 1:346–347
 constitutive equations 2:651
 dynamic multilevel methods 3:214–219
 geomaterials 2:545–549
 shakedown 2:316–320
 transonic potential flow 3:337–339
 viscoelastic fluid flow 3:494
 simulations
 adaptive wavelets 1:157–195
 aerodynamics 3:446–448
 civil aircraft 3:431
 renormalized scales 3:260–263
 turbulent flows 3:235–240
 see also direct . . .
 tests 1:180
 traces 3:91, 3:92, 3:115

object-order volume rendering 1:542
 objective functions 2:425, 2:654
 occludes 1:499–500, 1:577
 ODEs *see* ordinary differential equations
 Oldroyd-B fluids 3:482–488
 Oleinik's E-condition 1:444–445
 one-dimensional
 bars 2:676–677
 convergence 1:583–584
 finite volume methods 1:450–455
 hierarchical shape functions 1:120–121
 statically determinate structures 2:661–662
 wave propagation 3:531–533
 one-step methods 1:715–717, 2:182–183
 ONERA M6 wings 3:115, 3:116
 ONERA S1 Modane 3:429, 3:450–451
 open inventor toolkit 1:544
 OpenDX development environment 1:544–545
 OpenGL toolkit 1:544
 operation counts 1:604
 Operational Loads Survey rotor 3:104, 3:105
 operational quadrature method 1:704, 1:715, 1:717–719
 operators
 adaptivity 1:160, 1:165, 1:175–181
 advection-diffusion 3:18
 algebraic 3:30–32
 bending 1:214
 BGT 2:706
 biharmonic 1:34
 boundary integrals 1:342–344, 1:349, 1:351
 Clément 1:59–60, 1:66–68
 complementing 1:229
 consistent tangent 2:736–738
 cut 1:725, 1:727–734
 differential 1:238, 3:30–32
 element nodal interpolation 1:80–82
 equilibrium 2:294–295
 Euler 1:481–482, 1:483
 filtering 3:272–274
 five-point difference 1:9
 gradients 1:727–732
 heat potential 1:718–719
 Helmholtz 3:18
 hyperbolic 1:349, 1:376, 1:176, 1:177
 interpolation 1:80–82, 1:84, 1:58–60
 jumps 2:172

operators (*continued*)
 lifting 1:251–252
 linear 1:168–169, 1:191
 membranes 1:213
 nodal interpolation 1:80–82
 nonlinear 1:191–193
p-exact reconstruction 1:462–463
 residual error estimates 1:88
 scale separation 3:219
 Scott–Zhang 1:60, 1:66, 1:67
 sparse approximations 1:159, 1:175–181
 split methods 2:467, 2:590–591
 Skoldov–Poincaré 1:380–382, 1:384–386, 1:399
 superdissipative 3:241
 tangent stiffness 2:337–338, 2:339
 trace 1:351
 wavelets 1:160, 1:165, 1:168–169, 1:175–181
 optimization
 codes 3:412–413
 Dassault code SOUPLE 3:437–443
 elastic shakedown 2:301, 2:306, 2:310
 multigrid methods 1:578
 parameter identification 2:643–645
 projection-based interpolation 1:733
 theory 2:219, 3:437–443
 unit volume meshing 1:509–510
 ordinary differential equations (ODEs)
 adaptive computation 1:680, 1:699, 1:701
 first order 3:31–32
 shallow water equations 3:243–244, 3:249–250
 space-time formulations 3:31–32
 structural dynamics 2:170–189
 time step control 1:699, 1:701
 ordinary least squares 2:643
 orientation tensors 3:491–492
 orientations 1:322–323, 1:600–608, 2:426
 orthogonal iteration 1:569
 orthogonal polynomials 1:143–146
 orthogonality
 algebraic complements 1:730
 cohesive-zone models 2:352–353
 partitioning errors 2:424–425
 subsurface errors 2:424–425
 Orszag–Solomon flux 1:469
 out-of-core methods 1:543
 out-of-plane material models 2:536–537
 outflow boundary conditions 3:291
 overall reaction terminology 3:507–508
 overall testing processes 2:417–421
 overlap
 boundary integral equations 1:704–705
 incomplete LU factorization 1:563–564
 overlapping domains 2:575–599
 decomposition 1:617–620, 1:630–633
 nonoverlapping decomposition 1:61, 1:618–620, 1:633–644, 2:712
 overlay shakedown model 2:308–309
 overshoot 2:179–180, 2:181, 3:52–53

p convergence 1:126–131
p-exact reconstruction operators 1:462–463

p-finite element method 1:119–137
 convergence characteristics 1:126–131
 discretization 1:3
 elastoplasticity 1:133–135
 geometric nonlinearity 1:135–136
 implementation 1:120–126
 industrial applications 1:134–135, 1:136–137
 mapping 1:124–126
 nonlinear elastoplasticity 1:133–135
 one dimensional hierarchic shape functions 1:120–121
 performance characteristics 1:131–133
 thin domains, shells 1:219–229
 panel clustering 1:5–6, 1:597–615
 boundary element method 1:597–600, 2:731
 elasticity 2:731
 finite element method 1:597
 fully populated matrices 1:597–615
 hierarchical matrices 1:597, 1:607–615
 multiplication 1:600–601, 1:602–604, 1:607, 1:611
 panel methods 3:336–337, 3:580
 pantographs 3:5–6
 parabolic...
 differential equations 1:675–702, 3:341–342
 equations 1:18–28
 interpolation 1:125–126
 problems 1:3, 1:705–706, 1:710–719
 smoothing 1:688
see also adaptive computation
 parafolds 3:449–451
 parallel...
 edge-cracked borosilicate glass plates 2:39–40
 elimination 1:566
 iteration 2:711–712
 mesh adaptivity 1:516
 processing 3:476–477
 RKDG 3:112–113, 3:114
 visualization 1:543
 parameterizations
 bifurcations 1:669–673
 continuation 1:665–666
 local convergence theory 1:661–669
 Maxwell equations 1:731–732
 nonlinear equations 1:661–669
 parameters
 bifurcations 1:670–672
 derivatives 2:650–651
 identification 2:647–654
 panel clustering 1:603–604
 sensitivity 2:667–668
 parametric curves 1:504–505
 parametric surfaces 1:505–506, 1:479
 ParaView application 1:546
 Pareto fronts 3:393, 3:396
 partial differential equations (PDEs)
 discretization 1:15–17, 1:291–300
 elliptic 1:112–118
 geometric modeling 1:493
 meshfree methods 1:291–300
 weak solution 1:147–158
 partial pivoting 1:554–555
 partially...
 diagonal scaling preconditioner 1:592–593
 premixed flames 3:500–501, 3:503
 saturated cement pastes 2:517–520
 saturated consolidation 2:584–586

particles
 advection 1:534–535
 deformability 1:311–335
 distribution 1:286–287, 1:289
 fracture mechanics 2:391–392
 grid method 3:144–145
 integration form 1:293
 interactions 2:412
 motion 1:413–418
 redistribution 3:136, 3:140–141
 strength exchange (PSE) 3:137–139, 3:145–149
 traces 1:535–536
 velocity 1:417
 particulates 1:311–335, 2:418–420, 2:425–426
see also discrete element methods
 partitioning
 errors 2:424–425
 finite element spaces 1:79–80
 hierarchical matrices 1:608–609
 material responses 2:421–422, 2:425
 multifield problems 2:577–589
 panel clustering 1:601–602, 1:608–609
 partitions of unity (PU)
 discontinuous meshfree methods 1:301–303
 discrete failure models 2:365–369
 finite element method 1:291
 implicit error estimators 1:90–91
 meshfree methods 1:280, 1:291–292, 1:301–303
 moving least squares 1:291
 partial differential equations 1:291–292
 passenger aircraft 3:431–433, 3:434
 passive mechanical behavior 2:607–608
 passive myocardial tissue 2:624
 path following methods 2:151–153
 path functions 2:12
 path parameters 2:152–153
 PBX preconditioner 1:633
 PCG *see* preconditioned conjugate gradients
 PDE *see* partial differential equations
 PDEs *see* partial differential equations
 PDFs *see* probability distribution functions
 PEC *see* perfect electric conductors
 Pelet numbers 3:24
 pellet impacts 2:379
 penalty method
 contact discretizations 2:207
 contact mechanics 2:202, 2:203, 2:217
 discontinuous deformations 1:322–323, 1:328–329
 matrix form 2:209
 multibody contact forces 1:321–322
 partial differential equations 1:293, 1:296–297
 penetration...
 checks 2:215–216
 conditions 1:514, 1:528, 2:198–199, 2:201
 functions 2:199
 Peraire–Jameson flux 3:415
 perfect electric conductors (PEC) 1:726
 perfect incremental collapse 2:305
 perfectly matched layers (PML)
 acoustics 2:496, 2:710, 3:50–52
 eddy viscosity 3:50–52
 electromagnetics 3:50–52
 Maxwell equations 1:735
 perfectly stirred reactors (PSR) 3:513–514, 3:517
 perforated plates 2:526–529

performance evaluation 3:486–487
 performance functions 2:568, 2:569
 periodic...
 boundary conditions
 composite laminates 2:448, 2:449
 large eddy simulations 3:291
 multibody contact forces 1:323–324
 representative unit cells 2:443–445
 shallow water equations 3:242–245, 3:248, 3:249–250, 3:251–259
 flows 3:219–221, 3:235–240, 3:377–379
 forcing functions 1:27
 media theory 2:514, 2:515
 surface loads 2:592, 2:593, 2:594
 permeability property upscaling 2:516
 perturbation
 adaptive wavelets 1:158–159, 1:186–187
 Lagrangian 1:293, 1:322, 2:203
 postbuckling 2:147–149, 2:150
 saddle-point stability 1:254–256
 sensitivity 1:655–656
 stochastic finite elements 2:668–671
 Stokes equations 1:267
 Perzyna viscoplastic relation 2:235, 2:242–243
 Petrov–Galerkin (PG)
 direct linear algebraic solvers 1:556, 1:559
 elliptic boundary values 1:465–466
 ship hydrodynamics 3:583
 time finite element methods 2:186
 weak form 1:280, 1:291–292
 phase
 changes 2:584–586, 2:592–599
 errors 3:62–63
 stress-updating 1:429–431
 phenomenological theory 3:226–227
 photometric textures 1:491
 Picard iterations 1:594
 piecewise linear methods 1:661–662, 1:664
 piezoelectricity 2:758–759, 2:761–762, 2:763
 pinball technique 2:215
 Piola–Kirchhoff stress tensors 2:10, 2:66, 2:441–442, 2:444, 1:136
 Piola's transform 1:260–261
 pitchfork mapping 1:669–670
 pitching...
 airfoils 3:77–78
 cycles 3:380
 moment 3:451
 pivoting 1:570
 planar...
 boundary representation schemes 1:481–482
 Couette flow 3:489–490, 3:491
 domains 1:499
 flow past cylinders 3:487–488, 3:494–496
 polygons 1:481–482
 plane linear elasticity 1:377, 1:405–408
 plane-strain 2:15, 2:286–287, 2:504–506
 plane-stress 1:204–206, 2:15
 planetary model *see* periodic boundary conditions
 platforms 3:392–396
 plank equation 3:492
 plant physiology 1:685, 1:686
 plasma flows 3:81–85
 plastic...
 correctors 2:467
 crystals 2:280–283
 deformations 2:269–287

- plastic... (*continued*)
 elastic analysis 2:727–729, 2:738–740
 evolution equations 2:232–235, 2:241
 flow 2:274–275
 frictional tangential slip 2:206
 loading 2:235–236, 2:340–341
 metric 2:243
 multipliers 2:233, 2:315–316, 2:482, 2:738
 shakedown 2:324–325
 slip 2:273, 2:276–278, 2:206
 strain 2:466, 2:506
 tangent moduli 2:279–280
 volumetric strain 2:466
 work rates 2:487
 zone evolution 2:369
 plasticity
 bound limit theorems 2:549–551
 boundary integral equations 2:735–740
 composite laminates 2:451–452
 contact mechanics 2:205
 damage coupled models 2:340–341
 error indicators 2:486–487
 geomechanics 2:549–551, 2:552–553
 gradient-enhanced damage model 2:359–360
 shape sensitivity 2:740–743
see also crystal...; elastoplasticity; viscoplasticity
 plates 1:3
 asymptotic expansions 1:201–207, 1:215
 bending methods 1:85
 composite laminates 2:433–437
 coordinates 1:199–200, 1:202
 domains 1:199–200
 eigen frequencies 1:224
 finite elements 2:59, 2:79–83
 four-point bending 2:39–40
 fracture behavior 2:39–40
 with holes 2:326–329
 Kirchhoff–Love models 1:204, 1:208–209, 1:229, 2:433–435
 locking 1:221–222
 multiscale expansions 1:202–211
 shakedown 2:326–329
 shear locking 1:222–223
 shell theory 2:84–102
 transverse shear locking 2:122
see also thin-walled structures
 PLC *see* Portevin–le Chatelier
 ploughing 2:205
 ply level deformations 2:432, 2:453–458
 FML *see* perfectly matched layers
 Poincaré–Friedrich's inequality 1:75
 point...
 connections 1:500–501
 creation 1:500–501
 estimate method 2:568–569
 insertion 1:508
 moment closures 3:301
 placement 1:508, 1:510
 repositioning 1:510
 visualization techniques 1:528–529, 1:530–531
 vertices 3:131
 pointwise estimates 1:111
 Poisson solvers 3:144–145
 Poisson thickness locking 2:80, 2:97
 polarization 2:412
 pole-zero constitutive law 2:621, 2:624
 pollution 1:111–112, 2:696, 2:699–700, 2:701–702
 polycrystalline microstructures 2:267–287
 fluctuation fields 2:281
 heterogeneities 2:281, 2:282–283, 2:286–287
 polygonal patches 1:478, 1:481, 1:482
 polyhedral domains 1:73, 1:77–85, 1:107–109, 2:300–304
 polymer...
 melts 3:487–488, 3:490
 processing 3:481, 3:487–488
 stress 3:491–492
 polynomials
 algebraic 1:143–145
 approximations 2:700
 Bernstein 2:214
 Bézier 2:213–214
 bubble functions 3:20
 chaos expansion 2:671–676
 Chebyshev 1:143–144
 expansions 3:63–67, 1:143–145
 Fourier–Legendre 3:215–217
 Hermite 2:213–214, 2:671–676
 interpolation 1:604–605, 1:125–126
 Jacobi orthogonal 1:143–146
 Legendre 1:393, 2:704, 1:143–144
 spaces 1:285–286, 1:220–221
 spline 2:213
 pore...
 accessibility 2:518–519
 pressures 2:564–565
 size distributions 2:519
 water 2:561–562
 porous media 2:566–567, 2:584–586, 2:592–599
 Portevin–le Chatelier (PLC) band propagation 2:346, 2:349
 positional mismatches 1:547
 positive...
 coefficients 1:456–457, 1:463–464, 3:349–356
 streamline invariants (PSD) 3:415, 3:420–421
 symmetric semidefinite initial value problems 1:680, 1:681
 post-first ply failure 2:438
 postbuckling 2:147–149, 2:150
 postprocessing 1:526–527
 potential energy
 buckling 2:140–141, 2:147, 2:162
 material responses 2:422, 2:423
 Stokes equations 1:241
 potential flow methods 3:334–348, 3:580
 potential functions 1:322–323
 powder compaction 1:431–433
 power method 1:567–568, 1:569, 2:159
 power plant noise 3:444
 Prandtl–Batchelor flows 3:133
 Prandtl–Reuss equations 2:239
 prebuckling 2:147–149
 preconditioning
 conjugate gradients (CG) 1:642–643, 1:147
 domain decomposition 1:6, 1:617–644
 element-by-element 1:566
 hp-coupled BEM/FEM 1:390–393
 incomplete LU factorization 1:564–565, 1:566–567
 least squares coupled BEM/FEM 1:395–396
 linear algebraic solvers 1:555–556, 1:560–562
 minimum residual method 1:390
 multigrids 3:177
 Neumann–Neumann method 1:618
 nonoverlapping domain decomposition 1:618–619, 1:620, 1:633–644

- overlapping domain decomposition 1:617–618, 1:619–620, 1:630–633
 preconditioned conjugate gradients (PCG) 1:642–643, 1:147
 Schwarz theory 1:617
 time-stepping schemes 3:375–376
 predictors
 buckling 2:155
 corrector procedures 1:332–333, 2:153–156
 modified Lax–Wendroff flux 3:414
 prediction operators 1:165
 simulations 3:293
 preintegration 2:75, 2:98–102
 premixed combustion diagrams 3:518
 premixed flames 3:500–502
 preprocessing, visualization 1:526
 prescribed constrained shape methods 2:378–381
 pressure
 acoustics 2:685–686, 2:687, 2:697
 blood flow 3:528–540
 coefficients 3:101, 3:103, 3:435
 compressible fluid flow 3:122–123
 correction projections 3:68–69
 differences 3:204–205
 distribution
 adaptive wavelets 1:174
 aerodynamic shape optimization 3:380–381
 Mars Lander 3:327, 3:328
 time-stepping schemes 3:373–375, 3:378
 wing inverse design 3:396–397
 drag 3:443
 echo 3:516–517, 3:320–321
 gradient operators 3:583
 gradient projections 3:586
 head 2:565
 hydroelastic-clothing equations 2:690–691
 interpolation 1:262–266, 3:484
 reflection 3:316–317, 3:320–321
 sensitive point (PSP) 3:431–433
 stabilization 3:163
 structural-acoustics 2:685–686, 2:687
 volume relations 2:621
 presumed probability density function 3:520–521
 primal...
 augmented Lagrangian 2:260
 closest-point-projection 2:250–254, 2:262
 elastic shakedown 2:300
 meshes 3:101, 3:102
 partitioning 2:421–422
 Signorini-type interfaces 1:396–399
 primary states 2:141, 2:144–145
 primitives 1:483, 1:484
 principal curvatures 1:212
 principle...
 of frame indifference 2:13
 of maximum dissipation 2:237–239, 2:457, 2:466–467
 of Minimum Complementary Potential Energy 2:422
 of virtual work 2:16, 2:469–470
 prismatic elements 1:731
 probability...
 density functions 2:567–568, 3:520–521
 distribution functions (PDFs) 2:646
 of failure 2:568
 one homotopy methods 1:665
 probing 1:540, 1:547–548
 progress variables 3:511
 progressive damage modeling 2:451–458
 progressive failure 2:437–438
 Project SINUS, INRIS Sophia-Antipolis 3:473–477
 projections
 based interpolation 1:732–734
 Chorin scheme 3:168–170
 closest-point-projection 2:228–229, 2:244–246, 2:249–259, 2:262–263
 convective gradients 3:586
 elliptic 1:692–693
 filters 3:273
 Galerkin 1:207, 1:210
 interpolation 1:732–734
 Krylov 1:556–557
 L₂ 1:692–693, 1:142
 like operations 1:623, 1:631
 pressure 3:68–69, 3:586
 Ritz 2:32, 2:48–49
 schemes 3:168–170
 smoothers 3:179
 velocity-correction 3:67–68
 prolongations 1:580–582, 1:587–589, 1:591, 3:219
 proof techniques 1:269–276
 propagation
 blood flow 3:531–533
 instabilities 2:346–349
 Liders band 2:346, 2:348–349
 Portevin–Le Chatelier band 2:346, 2:349
 shear bands 2:348–349
 speeds 3:250–251
 waves 2:757–758, 3:531–533
see also crack...
 PSE *see* particle strength exchange
 pseudo arc lengths 1:666, 2:153, 2:154
 pseudo-overlap reordering, incomplete LU factorization 1:564
 pseudospectral collocation 3:485
 pseudospectral derivatives 1:142
 PSI *see* positive streamline invariants
 PSP *see* pressure sensitive point
 PSR *see* perfectly stirred reactors
 Ptolemaic cooling tower 2:533–538
 PU *see* partitions of unity
 public domain software 1:114
 pulse wave propagation 3:533
 punch compaction 1:431–432, 1:433
 pure mixing 3:508, 3:509–510
 Q-factors 1:652–653
 Q-orders 1:652–653
 Q_z-element 2:18–20
 QMR algorithm 1:559–560, 2:707
 QR factorizations 1:555, 1:558
 QR method 1:567, 1:568–571
 quadratic...
 kinematics 1:219
 macroscopic free energy 2:272
 strain energy function 2:466–467
 quadrature...
 errors 1:179–180, 1:603
 method 1:704, 1:715, 1:717–719
 points 2:113
 rules 3:108–109

quadrilateral elements
 finite element spaces 1:178, 1:181
 hanging nodes 2:38
 node-regular refinement 2:37–38
 Stokes 3:162–163
 Stokes equations methods 1:266
 thermal diffusion 1:260–261
 quadrilateral hierarchical shape functions 1:121–123
 quadtree-octree based methods 1:499–500
 quality meshing 1:502–510
 quasi...
 brittle fractures 2:350, 2:360–361, 2:363–364
 Eulerian conservation equations 1:419
 geostrophic flows 3:253–254
 interpolation 1:59–60, 1:66–68
 linear exact potential flow equation 3:342–343
 linear model 3:313
 sparsity 1:176–177
 static analysis 1:428, 2:160–164, 3:229–230
 triangular matrices 1:553
 quaternions 2:120
 QZ method 1:571

r-adaptive technique 1:422
 R-curves 2:396–399
 R-factors 1:652–653
 R-orders 1:652–653
 racing sailboats 3:596–601
 radial...
 basis functions (RBF) 1:300, 1:716
 return algorithm (RRA) 2:737–738
 set functions 2:709–710
 terms annihilation 2:705
 radius functions 1:485–490, 1:494
 radius-to-thickness ratio 2:87, 2:89–90, 2:92
 Rafale inlet design 3:428–429
 Rafale store releasing 3:429, 3:430
 rake angles 2:503, 2:504
 random...
 convection 3:315
 eddy 3:318
 fields discretization 2:663–667
 methods 3:293
 variables 2:658–661
 rank parameters 1:611
 Rankine sources 3:580
 Rankine yield criterion 2:535–536, 2:537
 Rankine–Hugoniot jump condition 1:440, 1:441
 RANS *see* Reynolds averaged Navier–Stokes; Reynolds averaged
 Numerical Simulations
 rapid trends 3:312–314
 rate dependent elastoplastic deformations 2:251–252, 2:254–255
 rate of plastic work 2:487
 Raviart–Thomas (RT) elements 1:259–261, 1:264–266, 1:271, 1:395
 ray tracing 1:484, 1:541–542
 Rayleigh quotients 1:567–568
 Rayleigh–Ritz method 1:348
 RBF *see* radial basis functions
 RBM *see* rigid bodies spring model
 reacting flow control 3:499–523
 reaction-diffusion equations 1:696–699, 1:700
 reactors, laminar flames 3:513–514
 real Schur form analysis 1:552
 real Shur decompositions 1:553, 1:570
 reciprocity 1:704, 2:758–763

recompression hierarchical matrices 1:613–614
 reconstruction schemes 3:354
 recovery estimators 1:93–95, 1:96–97
 rectangular elements 1:78–79, 1:108–107
 rectangular window functions 1:282
 recursive quadratic programming 2:318
 red-black ordering 1:563–564
 Red–Green–Blue refinement 1:102–103
 reduced...
 basis techniques 2:517–318
 gradient formulation 3:386–387
 integration 2:122
 reference configuration 2:64–65, 2:76, 2:81
 referential domains 1:414, 1:416, 1:417–418
 referential velocity 1:420
 REFINCOARSE 1:98, 1:99
 refined constitutive forming processes 2:481–484
 refinement
 flow field compression 1:170
 linearized elasticity 2:57–39
 rules 1:87, 1:100
 reflection invariance 3:275
 regional model adaptivity 2:44
 regular...
 data sets 1:529
 meshes 1:502–504
 triangulations 1:98, 1:100–103
 reinforced concrete 2:529–538
 Reissner–Mindlin (RM) model
 discretization 2:110–112
 plates 1:209–210, 1:224, 2:45, 2:110–112
 shell theory 2:86, 2:105–106, 2:110–112
 relative...
 errors 2:766, 2:767
 permeability 1:725
 permittivity 1:725
 relatively intact (RI) state 2:554
 relaxation
 elliptic 3:317, 3:318, 3:321
 function 2:753
 potential flow equation 3:344–345
 shutdown 2:315
 transonic potential flow 3:339
 transonic small-disturbance 3:340–341
 reliability
 adaptive computation 1:691, 3:197–198
 averaging error estimators 1:94–95
 error control 1:86
 geotechnical engineering 2:568–569
 indices 2:677
 methods 2:676–680
 residual error estimates 1:88
 remeshing 1:330–331, 2:363
 remodeling
 arterial walls 2:615–616
 heart wall mechanics 2:622
 rendering 1:527–528
 renormalization 3:259–263
 reordering unknowns 1:563–564
 representation
 formulas 1:355–358, 1:706
 geometric 2:378–394
 theorem of isotropic functions 2:444
 representative unit cells (RUC) 2:443–446

representative volume elements (RVE)
 composite laminates 2:432, 2:441–443, 2:445–451
 concrete mechanics 2:514–515, 2:520–521, 2:523
 homogenization methods 2:408, 2:423
 reproducibility 1:286, 1:287–288
 reproducing kernel particle method (RKPM) 1:284, 1:286
 reservoirs 2:692, 3:602, 3:603
 residuals
 approximations 1:187–195
 discontinuous Galerkin 3:93–94, 3:96, 3:103
 of the entropy inequality 1:447–448
 error estimators
 explicit 1:87–89, 1:96, 2:28–29
h version BEM/FEM 1:382, 1:383–384
 implicit 1:587–589, 2:29–31, 2:34–35, 2:50–52
 linearized elasticity 2:28–31
 a posteriori 1:86
 free bubbles 3:20–21, 3:35
 generalized minimal 1:558–559, 2:702, 3:565–566, 3:570
 Helmholtz equations 2:700–701
 Navier–Stokes code 3:417
 refinement 2:201–202
 strains 2:611–613
 stress 2:318, 2:613–616, 2:622
 resistance
 crack growth 2:396–399
 ship hydrodynamics 3:579, 3:580
 ship 2:273–274, 2:277
 resolution
 contact 1:318–321, 1:322–323
 direct numerical simulations 3:279–281
 large eddy simulations 3:281
 multiresolution 1:163–164, 1:480–481, 1:543
 rules 2:699–700
 switched schemes 3:351
 response
 functions 2:204
 nonlinear aeroelasticity 3:459–460
 statistics 2:669–671, 2:675–676
 restrained blocks 2:293–294, 2:321–323
 restriction
 multigrid methods 1:580, 1:587
 operators 3:219
 results visualization 1:526
 retaining wall structures 2:558–561
 retarded potentials 1:706, 1:707, 1:711, 1:712, 1:713
 reurn mapping 2:444–250, 2:219, 2:467–468
 Reuss fields 2:409, 2:411–412
 reverse transforms 1:172–174
 Reynolds averaged Navier–Stokes (RANS) equations
 computational aerodynamics development 3:329
 Dassault Aviation solver 3:423–426
 multigrid time-stepping 3:374
 ship hydrodynamics 3:580, 3:593
 turbulence closure 3:301–303, 3:305, 3:318–322
 turbulent flames 3:515–517, 3:520
 Reynolds averaged Numerical Simulations (RANS) 3:270
 Reynolds averaged turbulence modeling 3:318–322, 3:418–423
 Reynolds averaged velocity 3:301
 Reynolds numbers 3:2, 3:184, 3:210–213
 Reynolds stress
 Second Moment Closure 3:315, 3:316
 ship hydrodynamics 3:593–594
 tensors 3:274, 3:302, 3:303
 transport 3:320
 turbulence closure 3:302–303, 3:315, 3:316, 3:319–322

Reynolds transport theorem 1:419
 rheoelectric analog computers 3:409
 rheology 3:481, 3:487–488, 3:489
 RI *see* relatively intact
 rib scales 2:529–530
 ribbons 1:536
 Riccati equation 1:615
 Richardson extrapolation 1:11–12
 Richardson iteration 1:555–556, 1:580
 Riemann solvers 3:577, 3:598–100
 Riemannian metrics 1:509–510, 2:8–9
 Riesz basis 1:162–163, 1:166–167
 right-preconditioning 1:561
 rigid...
 bodies
 collisions 1:314–335
 fluid interactions 1:424, 1:425–426
 nonlinear structural dynamics 2:185–186, 2:187
 search strategies 2:216
 spring model (RBSM) 1:333–334
 cube within water 3:604–607
 models 2:495–496
 motionless cavities 2:691
 ships 3:601, 3:602–605
 Rioja de España 3:596–601
 Ritz...
 Galerkin methods 1:73, 1:74–77, 1:556, 1:557–558
 projections 2:32, 2:48–49
 values 1:553, 1:561, 1:572
 vectors 1:553, 1:572
 RI-matrices 1:609–610
 RKDG *see* Runge–Kutta discontinuous Galerkin
 RKPM *see* reproducing kernel particle method
 RM *see* Reissner–Mindlin
 RMD approximation 1:284
 robustness 1:132–133
 rock mechanics 2:543–569
 rods 2:757
 Roe flux 1:469
 rolling ball fillets 1:133
 rotation
 invariance 1:733–734, 3:275
 magnitude 2:85, 2:87–91
 oscillations 1:425–426
 parameterization 2:77, 2:116–120
 propellers 3:571, 3:572
 representative volume element 2:441
 rotated differencing 3:342–343
 rotated Stokes elements 3:162
 shallow water equations 3:251
 tensors 2:77, 2:117–120
 turbulence closure 3:314, 3:320
 vectors 2:118–119
 water mills 3:603
 Rothe Method 3:166–167
 rough inverse hierarchical matrices 1:615
 roundoff errors 1:656–657
 row-orientation 1:600–606
 RRA *see* radial return algorithm
 RST equations 3:311, 3:317–318
 RT *see* Raviart–Thomas
 RUC *see* representative unit cells
 Runge–Kutta
 discontinuous Galerkin method (RKDG) 3:97, 3:104–117
 explicit time-stepping 3:567–568, 3:572

Runge-Kutta (*continued*)
 time integration 2:176, 2:178
 visualization algorithms 1:534-535
 RVE *see* representative volume elements

S-A *see* Spälarst-Allmaras model
 S1 Modane parafol model 3:450-451
 S1 Modane stress release testing 3:429
 saddle points 1:246-257, 1:182-183
 safety assessments 2:291-331
 safety factors 2:305, 2:551, 2:568
 St. Venant-Kirchhoff-type material law 2:566, 2:584
 sampling 2:417-421
 sandstone soil consolidation 2:583-584
 saturation 2:31
 assumption 1:91-93, 1:103
 cement pastes 2:517-520
 consolidation 2:584-589
 edge 1:508
 geomechanics 2:553-554
 materials 2:553-554, 2:566-567
 soils 2:589-592

scalar
 advective-diffusive equations 3:32-37
 conservation laws 1:439-450, 3:97-98, 3:100-108, 3:349-351
 fields 1:530, 1:531-533, 2:741
 hyperbolic equations 1:28-36
 parameters 1:670-672
 turbulence closure 3:303-311, 3:320
 variable models 3:303-311
 scale effects 2:360-362
 scale separations
 discrete Navier-Stokes equations 3:219-228
 dynamic multilevel methods 3:207-264
 Navier-Stokes equations 3:213-214, 3:219-228
 renormalization theory 3:260-263
 shallow water equations 3:248-259
 turbulence 3:40, 3:45-46

scaled...
 boundary element method 2:547
 moving least squares 1:288-289
 norms 3:121-122
 scales of observation 2:513-539
 scaling updates 2:278-279
 scalp 1:520-521
 scanning-electron microscopy (SEM) 2:518-520
 Schmid stress 2:272-273, 2:276-277, 2:279
 Schur complement
 adaptive wavelets 1:186
 eigenvalues 1:570
 error indicator 1:384-387
 incompressible viscous flows 3:177
 preconditioning 1:618, 1:635-636
 Schur matrix forms 1:553, 1:569, 1:570
 Schwarz theory 1:617-644
 scientific visualization 1:525-526
 SCIRun development environment 1:545
 Scott-Zhang operators 1:60, 1:66, 1:67
 scripts 1:492
 SD-DF *see* streamline diffusion discontinuous Galerkin
 seabed geomechanics 2:555-558, 2:559
 search costs 3:389-390
 search procedures 1:317-321, 1:535, 2:196, 2:214-216

second...
 kind integral equations 1:594
 law of thermodynamics 2:12-13
 moment closure (SMC) 3:311-318, 3:320
 moment transport 3:311-318
 type tetrahedral elements 1:728
 second-order...
 elliptic boundary values 1:107-111
 ellipticity 3:91, 3:115-120
 hyperbolic equations 1:28, 1:30-31
 linear transmission 1:394-395
 monotone schemes 3:420
 predictors 2:155
 problems 1:83-85
 reliability method 2:678
 Sedov-type explosions 3:115, 3:116
 seepage 2:565
 seismic behavior 2:592, 2:593, 2:594
 self contact 2:216
 Self-Consistent method 2:412
 SEM *see* scanning-electron microscopy; spectral element method
 semi-implicit time integration 3:245-246, 3:252, 3:254-256
 semi-Lagrangian time integration 3:69
 semidiscrete...
 finite elements 3:565-566
 finite volume methods 1:443
 multiscale formulation 3:54-55
 space hierarchies 1:218
 semidiscretization 3:462-464
 seminorms 1:74-75
 semipolynomial spaces 1:204
 semisubmerged rotating water mills 3:603
 SEMMT 3:554
 sensitivity
 buckling 2:149-150
 elastic deformation 2:275-276
 imperfection 2:149-150, 2:349
 parameterized nonlinear equations 1:667-669
 spherical shells 1:224-226, 1:229
 stochastic finite elements 2:667-668
 separations
 cork of separation 2:350-351
 flow 3:320
 separability-of-scales 2:514
 shallow water equations 3:248-251
see also scale...
 sequential quadratic programming 2:216, 2:218-219
 Serendipity elements 1:78, 1:81
 Serial Staggered Procedure 3:473
 seven-parameter shell model 2:84, 2:90, 2:97-98
 SGBEM *see* symmetric Galerkin boundary element method
 shadows 1:541
 shakedown
 algorithms 2:312-321
 alternating plasticity 2:304
 applications 2:321-330
 classical 2:296-299
 elastic simulations 2:319
 extremum principles 2:299-301
 hardening 2:307-310
 incremental collapse 2:304-307
 kinematics 2:294-302, 2:305, 2:318-319
 loading-unloading conditions 2:293-294, 2:296-297, 2:300-304, 2:321-326
 numerical procedures 2:316-320

performed plates 2:326-329
 plastic 2:304
 plates with holes 2:326-329
 safety assessments 2:291-331
 square plates 2:326-327
 temperature 2:307-310
 thermomechanical loadings 2:293-294, 2:321-323
 tubes 2:323-326
 shallow...
 foundations 2:555-558, 2:559
 shells 1:215-217
 water equations 3:240-259
 shape...
 functions
 continuous discretization 2:664
 degenerating solid elements 2:79-80
 derivative evaluation 1:283-291
 Maxwell equations 1:732-733
 mixed finite element methods 1:238
 modifications 1:293
 one dimensional hierarchical 1:120-121
 partition-of-unity 2:365-369
 optimization 3:79-400, 3:437-443
 regular elements 1:57-58, 1:68
 sensitivity analysis 2:740-743
 sharp cutoff filters 3:273-274
 shear
 angles 2:503, 2:505
 bands
 cohesive-zone 2:351-355
 Liders propagation 2:348-349
 meshfree methods
 Portevin-Le Châtelier propagation 2:349
 rake angles 2:503-504
 correction factors 1:209-210, 2:97
 crystal plasticity 2:283-284
 energy 1:209
 flows 3:294
 geomechanics 2:557, 2:558
 layer instability 3:123
 locking 1:222-223
 material responses 2:418
 moduli 2:409, 2:418, 2:419
 slip mesh update method (SSMUM) 3:555, 3:571, 3:572
 stiffness 2:339
 strain 1:206, 2:536-537
 stress
 arbitrary Lagrangian-Eulerian 1:433
 blood flow 3:535, 3:536-537
 concrete mechanics 2:536-537
 transport (SST) 3:508-509, 3:511, 3:520
 turbulence closure 3:310, 3:311
 abedding frequencies 3:321
 sheet metal stamping 1:518, 1:519
 shells 1:3, 1:218-219
 asymptotic expansions 1:201-202, 1:211-218
 coordinates 1:199-200
 director definition 2:113-116
 domains 1:199-200
 eigen frequencies 1:224-229
 finite elements 2:59, 2:79-83, 2:104-128
 hierarchical models 1:223-224
 higher-order models 2:106-107
 intersections 1:133
 Kirchhoff-Love models 1:211-218, 2:103-105

layer-wise models 2:106-107
 limiting models 1:211-218
 locking 1:221-223
 membrane theory 2:102-103
 multiscale expansions 1:211-218
 Reissner-Mindlin models 2:105-106
 shifters 2:73-74, 2:94-95, 2:98-102
 structural behavior 2:68-70
 theory derivation 2:72-102
see also thin-walled structures
 shift strategies 1:570
 Shift Theorem 1:76-77
 shifter tensors 2:73-74, 2:94-95, 2:98-102
 ship hydrodynamics 3:379-407
 characteristic length parameters 3:592-593
 finite calculus 3:581, 3:583-587
 finite element discretization 3:587-589
 fluid dynamics 3:3
 fluid-chip interactions 3:589-590
 Lagrangian fluid flow 3:581, 3:591-592
 mesh node updating 3:590-591
 Navier-Stokes equations 3:581-583
 ship motion 3:389-390
 transient atom flow 3:591
 turbulence 3:593-594
 viscosity 3:592
 shock
 capturing
 compressible flows 3:551-553
 computational aerodynamics 3:329, 3:460
 discontinuous Galerkin methods 3:97, 3:98-104
 Euler equations 3:348-359
 fluid flow discretization 3:363-364
 Navier-Stokes equations 3:348-359
 nonoscillatory 3:348, 3:349-356
 RKDG 3:111-112
 complex geometry 3:346-348
 direct numerical simulations 3:281, 3:294
 free flows 3:346
 large eddy simulations 3:281, 3:295
 point difference equations 3:341, 3:342
 RKDG 3:111-112
 waves 3:460
 Shortley-Weller approximation 1:13
 shotcrete 2:514, 2:516-529
 shrinkage strains 2:520-521
 throats 3:5-6
 tide modes 1:122, 1:123
 side-grooved specimens 2:398
 sign conditions 3:107
 significant coefficient predictions 1:189, 1:191-193
 Signorini-type interfaces 1:376-377, 1:396-403
 silicon carbide-carbon 2:345, 2:351
 Silver-Müller condition 1:735
 simple mechanism of incremental collapse (SMIC) 2:304-307, 2:323, 2:325
 simulation
 crack growth 2:377
 flow 3:570-574
 implosion 3:115, 3:117
 platforms 3:473-477
see also direct numerical...; large eddy...; numerical...
 single...
 column supports 2:44-46
 crystal plasticity 2:274-280

single... (continued)
 edged notched beams 2:352, 2:354–355, 2:363
 field time-discontinuous Galerkin methods 2:176
 layer potentials 1:599
 phase nonreactive fluids 3:269–296
 step chemical reactions 3:510
 singular integrals 2:725
 singular value decompositions (SVD) 1:555
 singularity prevention 3:305
 six-parameter shell model 2:84, 2:87–88, 2:97
 size effects 2:360–362, 2:363
 skeletons 1:536
 skew-symmetric convection 1:149
 skin friction 3:307, 3:315, 3:321–322
 slave nodes 2:212–213
 sliding 2:197, 2:199–202, 2:206, 2:210–214, 2:220
 SLIP *see* symmetric limited positive slip
 slip
 concrete mechanics 2:531
 contact mechanics 2:201–202, 2:205, 2:206
 resistance 2:273–274
 system updating 2:278–280
 updating 2:278–280
 Slobodetskii norms 1:362
 slope limiters
 discontinuous Galerkin 3:97, 3:98–100, 3:104
 multidimensional finite volumes 1:459–460
 RKDG 3:107–108, 3:109
 sloshing modes 2:691
 slow terms 3:312, 3:313
 Smagorinsky model
 eddy viscosity 3:40–41, 3:43–44
 large eddy simulations 3:424–425
 ship hydrodynamics 3:598–599
 subgrid-scale modeling 3:285–286
 viscosity 3:241
 small...
 eddies 3:259–263
 scale decompositions 3:219–228
 scale separations 3:248–251, 3:260–263
 strains 2:208–210, 2:467–468, 2:647–649
 SMC *see* second moment closure
 smeared cracks 2:394–399
 smeared format 2:351
 SMC *see* simple mechanism of incremental collapse
 smooth
 average fields 3:301
 filices 3:273–274
 particle hydrodynamics (SPH) 1:280, 1:281–285, 1:291
 solutions 1:129
 surfaces 2:213–214
 variational multiscale method 3:13–15
 smoothing effect 1:581–582
 smoothing iterations 1:580
 snap back points 2:149
 snap-backing 2:148–149, 2:159–164
 snap-through 2:70, 2:148–149, 2:159–164
 snapping 2:148–149, 2:159–164
 Sobolev 3:389
 gradient 3:389
 index 1:366–371
 inner products 3:388
 scale 1:177
 spaces
 adaptive wavelets 1:166, 1:188
 boundary integral equations 1:710–712

convergence optimality 1:362–363
 finite element methods 1:104
 linear elliptic boundary values 1:74–75
 symmetric coupled BEM/FEM 1:378
 Variational Alternating Schwarz Method 1:620
 soft biological tissue 2:605–629
 soft limiters 3:553
 softening
 concrete mechanics 2:531–532, 2:536
 geomaterials 2:553, 2:555–556
 nonlinear 2:451–452
 software
 architecture 1:475–476
 cardiovascular surgery 3:538–539
 finite element methods 1:98, 1:113–114
 parabolic differential equations 1:702
 shells 1:219
 soil
 behavior 2:577–589
 consolidation 2:561–567
 dynamics 2:589–592
 mechanics 2:543–549
 skeletons 2:561–562, 2:564–565
 structure interactions 2:598–561
 solid...
 boundaries 3:145–149
 extension mesh moving techniques (SEMMT) 3:554
 textures 1:491–492
 wall boundary conditions 3:291–292
 solids
 constitutive models 2:1–3
 deformability 1:527
 elastoplastic deformations 2:227–264
 instability 2:1–3
 introductory survey 2:1–3
 mechanics 1:426, 1:428–433, 1:497–521
 multifield problems 2:1–3
 multiscale modeling 2:1–3
 nonlinearity 2:1–3
 processing 2:1–3
 structures 2:1–1–38
 viscoplastic deformations 2:227–264
 within water 3:604–607
 solution interpolation 1:516
 solution regularity 3:157–159
 solvability 1:246–247
 solvers 1:4–5, 3:120, 3:408–412
 Sommerfeld radiation condition 2:702, 2:708
 sound pressure 3:6–7
 SOUPLE 3:437–443
 space splitting 1:623–633
 space vehicles 3:412, 3:433–435
 space-periodic boundary conditions 3:209
 space-time
 averages computability 3:199–201
 boundary integral equations 1:703–713
 contact technique (STCT) 3:557–558
 discrete maximum principle 1:445–446
 finite elements 2:171–173, 3:27–28, 3:565–566
 Galerkin finite elements 1:676, 1:684–686, 1:689–690
 incompressible Navier-Stokes equation 3:44–45
 multiscale method 3:27–32
 stabilized 3:549, 3:555–557, 3:571–574
 structural dynamics 2:171–173
 variational multiscale method 3:27–32
 wave equation 1:706–707, 1:711–713

Spalart-Allmaras model (S-A) 3:519, 3:520
 spar-load distributions 3:380
 sparsity
 adaptive wavelets 1:158–159, 1:175–181
 approximations 1:158–159
 matrices 1:555
 patterns 1:555
 spatial...
 discretization
 averaging step functions 2:664
 Euler codes 3:413–414
 incompressible flows 3:69, 3:160–166
 incompressible Navier-Stokes equations 3:160–166
 RKDG 3:105–106
 shallow water equations 3:242–246
 thin-walled structures 2:71–72
 viscoelastic fluid flows 3:484–485, 3:494
 viscous flows 3:160–166
 domains 1:414, 1:416, 1:417–418
 location 1:540
 occupancy enumeration 1:476–477, 1:484–485, 1:532, 1:533
 periodic boundary conditions 3:209
 periodic flows 3:210–212, 3:555–557, 3:573
 representation 1:528–530
 search strategies 2:215
 semidiscretization 2:71–72
 tangent modulus 2:468–469
 turbulent flows 3:225–228
 species equation 3:505, 3:514
 specific...
 entropy 2:12
 strain-energy functions 2:13, 2:14–16
 time integration 3:252–259
 spectral...
 approximation 3:220–221, 3:242–245, 3:248–259
 collocation flows 3:485
 convergence 1:142
 eddy viscosity 3:50
 element method (SEM) 3:1, 3:81–85, 1:150–152
 equivalence estimates 1:627–630
hp elements 3:61–88
 methods 1:13, 1:141–154
 advection equations 1:148–150
 algebraic expansions 1:143–146
 conservation laws 1:148–150
 Fourier methods 1:141–142
 Jacobi orthogonal polynomials 1:143–146
 mortar method 1:152–154
 Navier-Stokes equations 1:146–148
 orthogonal polynomials 1:143–146
 polynomial expansions 1:143–145
 spectral elements 1:150–152
 Stokes equations 1:146–148
 trigonometric expansions 1:141–142
 radius 1:332–333
 random fields 2:663–667
 stochastic finite elements 2:671–676
 viscosity 1:150
 SPH *see* smooth particle hydrodynamics
 spheres falling in liquid filled tubes 3:571–573
 spherical...
 Hankel functions 2:704–706
 harmonics 2:704
 window functions 1:282
 spiral bevel gears 2:388–390

spine
 polynomials 2:213
 surfaces 1:479–480
 wavelets 1:480–481
 window functions 1:281, 1:283–284
 split stress-updating 1:429–431
 splitter plates 3:425–426
 splitting 1:263–264, 3:168–170
 springs 2:668–669
 SQP *see* successive quadratic programming
 square cylinder drag 3:191–192, 3:193–194, 3:197–198
 square plates 2:326–327
 SRVE *see* statistically representative volume elements
 SSMUM *see* shear slip mesh update method
 SSP *see* strong stability preserving
 SSP-RK *see* strong stability preserving Runge-Kutta
 SST *see* shear stress transport
 stability
 advective-diffusive equations 3:33–36
 buckling 2:142–144, 2:145–149, 2:150–164
 conservation laws 1:445–446, 1:467–468
 damage mechanics 2:342–344
 discontinuous Galerkin methods 3:92–93, 3:95–97, 3:106–107, 3:119
 elastic-plastic crystals 2:280
 envelopes 3:489
 factors
 adaptive computation 1:675–679, 1:683, 1:686–689, 1:691–693
 laminar flow 3:204, 3:205
 parabolic differential equations 1:675–679, 1:683, 1:686–689, 1:691–693
 Galerkin boundary elements methods 1:365–366
 geomaterials 2:551
 geomechanical engineering 2:567–569
 heterogeneous microstructures 2:282–283
 least squares constitutive equations 2:645–646
 mixed finite element methods 1:246–257
 nonlinear aeroelasticity 3:459–460
 parameter identification 2:641
 polycrystallines 2:282–283
 preservation 1:21–22
 shallow water equations 3:244–245, 3:246
 Sobolev index 1:367–368
 Stokes equations 1:266–268
 thin-walled structures 2:126
 time integration 2:178–179, 2:188–189
 weights 1:677
 stabilization
 adaptive wavelets 1:159–161
 arbitrary Lagrangian-Eulerian 1:414
 compressible flows 3:551–553
 fluid dynamics 3:548–550, 3:551–553
 incompressible flows 3:549–550
 stabilized
 integrals 3:585
 Maxwell equations 1:727
 methods 3:1
 advective-diffusive equations 3:32–40
 concept 3:22–26
 Dirichlet-to-Neumann formulation 3:8–11
 Galerkin 3:183
 multiscale methods 3:5–55
 space-time formulations 3:30–32, 3:549, 3:555–557, 3:571–574
 stable flames 3:500–501, 3:503–504
 staggered meshes 3:223–225

staggered partition solution procedures 2:580–582, 2:583
 stagnation enthalpy 3:358
 standard...
 staggered partitioning 2:580–582
 stress updating 2:276–277
 time step control 1:699, 1:701
 wavelet representation 1:168
 Stanton number isolines 3:434
 starting conditions 2:180–181
 state...
 equations 3:417–418, 3:439, 3:506–507
 functions 2:12
 variables 2:270–271
 static...
 condensation 1:263, 2:96
 constraint 2:339–340
 determinate structures 2:661–663
 equilibrium states 2:140
 finite elements 2:126, 2:303–304
 instabilities 2:341–346
 linear elasticity 1:622–623
 pressure 2:685, 2:687, 2:690
 response 2:691
 stationary...
 discrete shocks analysis 3:357–358
 heat conduction 1:622
 problems 1:158, 3:201–205
 statistical...
 closure 3:301
 linear elastic continua 2:670–671, 2:675–676
 modeling 3:519, 3:520–521
 moments 2:670–671
 noise modeling 3:444–445, 3:446–448
 statistically representative volume elements (SRVE) 2:423
 STCT *see* space-time contact technique
 steady...
 advective-diffusive equations 3:32–36
 flows 3:358–359, 3:481, 3:482–488
 one-dimensional tools 3:514
 processes 1:428
 state 3:459
 viscoelastic fluid flow 3:481, 3:482–488
 stealth aircraft 3:430–431
 steel
 autogenous 2:649, 2:650
 axisymmetric necking 2:651–652, 2:653–654
 concrete interaction 2:513–514, 2:529–538
 pellet impacts 2:379
 sheets 2:474–476
 see also metals
 steepest edge active set scheme 2:550
 Steger–Warming flux vector splitting 1:469–470
 Steklov–Poincaré operators 1:380–382, 1:384–386, 1:399
 step function discretizations 2:664
 step relaxation 2:515
 stepping procedures 2:151–156
 stick 2:199–202, 2:204, 2:205, 2:220
 stiff initial value problems 1:680–686, 1:687
 stiffness
 backfill soil 2:538–561
 degradation 2:553
 equation inversion 2:671–673
 matrices
 cohesive-zone models 2:353
 concrete mechanics 2:530

discontinuous deformations 1:328, 1:332
 eigenvalues 1:571
 elastoplasticity 2:745
 hierarchical 1:611–612
 MATLAB 1:113–114
 panel clustering 1:611–612
 soil skeletons 2:565
 thin-walled structures 2:125
 reduction schemes 2:453, 2:454–456
 stochastic
 constitutive equations 2:644–645, 2:646–647, 3:490–491,
 3:493–494
 finite element methods 2:657–680
 Karhunen–Loève expansion 2:665–667, 2:671–676
 perturbation 2:668–671
 random fields representation 2:663–667
 random variables 2:658–661
 reliability methods 2:676–680
 spectral formulations 2:663–667, 2:671–676
 statically determinate structures 2:661–663
 variable sensitivity 2:657–668
 geomechanics 2:567–569
 optimization 2:644–645
 reacting flow stoichiometry 3:507–508
 response 2:673–675
 stoichiometry, reacting flows 3:507–508
 Stokes elements 1:182–183, 3:161–163, 3:164
 Stokes equations
 inf–condition 1:269–276
 mixed finite element methods 1:240–241, 1:262–268
 saddle-point stability 1:249–251
 spectral methods 1:146–148
 see also Navier–Stokes equations
 stopping criterion 1:86
 store releasing, military aircraft 3:429, 3:430
 stored energy 2:230–231, 2:234
 strain
 arterial walls 2:611
 bounds 2:311
 composite laminates 2:441–442, 2:446, 2:455, 2:457–458
 concrete mechanics 2:520–524, 2:526–529, 2:536
 constitutive equations 2:647–654
 contact discretizations 2:208–214
 coupled damage-plasticity 2:340–341
 degenerating solid elements 2:81–83
 direct measures 2:77–78
 elasticity 1:245–246, 2:9–10
 elastoplasticity 2:235–237
 energy
 clamped hemispherical shells 2:69–70
 compressible materials 2:13–16
 elasticity constitutive tensors 2:13
 hyperelastic 2:466–467
 isotropic compressible materials 2:14–16
 principle of frame indifference 2:13
 shell theory 2:87–89
 through-the-thickness integration 2:94
 enhanced 1:245–246, 1:267–268, 2:93
 fields 1:326–329, 2:399–400, 2:457–458
 geomechanics 2:553, 2:555–558, 2:559–2:566
 hoop 2:87–89
 incompressible elasticity 1:244–246
 interior points 2:735–736
 kinematics 2:9–10
 localization 2:521–522, 2:553

logarithmic measures 2:464–465
 material responses 2:410
 measures 2:66–67, 2:77–78, 2:464–465
 plates 1:200, 1:206
 rate 2:506, 3:259–310
 representation 2:722–723
 return maps 2:467–468
 shell theory 2:87–93, 2:95–98
 softening 2:342, 2:344, 2:345–346, 2:347
 soil consolidation 2:566
 Stokes equations 1:267–268
 tensors
 Almansi 2:9–10, 1:136
 boundary integral equations 2:720
 clamped elliptic shells 1:213, 1:214
 concrete mechanics 2:522–524
 continuum damage mechanics 2:455
 degenerating solid elements 2:82
 effective... 2:455
 elastic body deformations 2:9–10
 Euler 1:136
 forming process modeling 2:463–464
 Green–Lagrangian 2:66–67, 2:73–74, 2:91, 2:441–442
 Hencky 2:9
 p finite element method 1:136
 shells 1:216, 2:91–93, 2:94–95
 statistical moments 2:670–671
 visualization algorithms 1:537–538
 thin-walled structures 2:66–67
 three-dimensional continuum 2:74
 transverse shear locking 2:122–123
 see also finite...
 Surug lemma 1:144
 strooklines 1:535–536
 stream ribbons 1:536
 stream surfaces 1:536
 streaming data 1:543
 streamline diffusion discontinuous Galerkin (SD–DD) 1:449–450, 3:97
 streamline-upwind/Petrov–Galerkin (SUPG) method
 advective-diffusive equations 3:34–35, 3:37
 compressible flows 3:551–553
 ship hydrodynamics 3:583, 3:592
 stabilized methods 3:22–23
 viscoelastic fluid flows 3:482–484, 3:489–490, 3:495
 streamlines 1:536
 curvatures 3:320–321
 diffusion 3:187, 3:188–189
 visualization algorithms 1:535–536, 1:538, 1:539
 stress
 arterial walls 2:611
 collocation 2:727
 composite laminates 2:432, 2:439–446, 2:454, 2:456
 concentrations 2:432, 2:439
 concrete mechanics 2:522–531, 2:535–537
 elasticity 1:244–246, 2:10
 elastoplasticity 2:230–231, 2:235–237, 2:238–239
 fields 1:327–328
 incompressible elasticity 1:244–246
 integration 2:467–468, 2:493–495
 intensity factors 2:734
 interior points 2:735–736
 low-order elements 2:477–478
 material responses 2:410
 plates 1:200, 1:205–206
 power 2:230–231

principle 2:561
 representation 2:722–723
 resultants 2:75, 2:78–79
 scaling 2:315
 shells 1:200, 2:75, 2:78–79, 2:95–98
 state 2:238–239
 strain 1:377, 2:336, 2:555–559
 Stress Check software 1:219
 tensors
 acoustic field equations 2:696
 Biot 2:10
 boundary integral equations 2:720
 Cauchy 1:418, 2:66, 2:456, 1:136
 concrete mechanics 2:522–524
 constitutive equations 2:650–651
 continuum damage mechanics 2:454
 cross 3:274
 damage mechanics 2:336
 effective 2:454
 elasticity balance equations 2:10
 elastodynamics 2:767
 Kirchhoff 2:10, 2:14, 2:230–231
 Piola–Kirchhoff 2:10, 2:66, 2:441–442, 2:444, 1:136
 statistical moments 2:670–671
 thin-walled structures 2:66–67
 three-dimensional continuum 2:73–74
 turbulence closure 3:311
 visualization algorithms 1:537–538
 turbulence closure 3:310–311
 updating 1:428, 1:429–431, 2:274–280
 stretched flames 3:512–513
 stretching 1:202, 2:80, 2:491, 3:130–131
 strings 1:9–10
 strip stretching 2:491
 strong...
 boundary value problems 2:16–17
 discontinuity kinematics 2:351–355
 stability factors 1:675–679, 1:683, 1:686–689, 1:691–693
 stability preserving Runge–Kutta (SSP-RK) 3:104–105, 3:106–107
 stability preserving (SSP) time integration 1:464–465
 Strouhal number 3:84–85
 structural...
 acoustics 2:683, 2:684–689
 dynamics 2:169–189
 acceleration equations 2:173
 elastodynamic equations 2:170–171
 equation formulation 2:170–174
 momentum equations 2:173–174
 multibody equations 2:174
 nonlinear 2:184–187
 ordinary differential equations 2:170–189
 space-time equations 2:171–173
 time integration 2:175–189
 transient analysis 2:169–189
 elasticity 2:40–42
 material responses 2:407–427
 mechanics 2:40–42
 scales 2:530–538
 stiffness 2:70
 subgrid-scale modeling 3:284, 3:287, 3:290–291
 thin-walled structures 2:70–107
 structure
 arterial wall 2:606–607
 function subgrid-scale modeling 3:286
 heart wall 2:618–619

structure (continued)
 icelastic material constitutive equations 2:639
 ligaments 2:625–626
 meshes 1:456–457, 3:359–360
 soil 2:554–555
 subcycling methods 2:188
 subdomain solvers 1:636–638
 subfilter scales 3:234–290
 subgrid-scale models
 direct numerical simulations 3:184–185, 3:198–199
 generalized Galerkin 3:184–185, 3:198–199
 large eddy simulations 3:184–185, 3:198–199, 3:270–271, 3:272–274, 3:423–424
 ship hydrodynamics method 3:583
 Smagorinsky eddy viscosity 3:43
 space-time formulations 3:27–32
 stabilized methods 3:22–26
 turbulence 3:46–47, 3:284–291
 subgrid stress 3:32–33
 subparametric mapping 1:58
 subsonic
 flows 3:101–104, 3:334–337, 3:425–426
 jets 3:446–448
 linearized potential flow 3:334–337
 mixing enhancement 3:425–426
 points 3:342–343
 subspaces
 bilinear forms 1:623–633
 interaction lemma 1:629–630
 iteration 1:590
 sequences 1:207
 subspace errors 2:424–425
 substructuring 2:686–689
 subtraction 1:611
 successive quadratic programming (SQP) methods 2:218–219
 sufficient conditions 2:297–299, 2:306–307
 supercomputing 3:411–412
 superconvergence 1:112–113, 2:32–33
 supercritical airfoils 3:460
 superdisjunctive operators 3:241
 superquadratics 1:520, 1:531
 supersonic
 flight 3:390
 flows 3:101, 3:102–104, 3:425–426
 jets 3:446–448
 mixing enhancement 3:425–426
 points 3:342–343
 zones 3:344–345
 SUPG *see* streamline upwind Petrov-Galerkin
 support stencil reconstructions 1:462–463
 surface...
 currents 1:726
 curvature 3:320
 domains 1:499
 fatigue wear 2:206
 meshing
 adaptivity 1:514–516
 advancing-front 1:501
 Delaunay-type 1:502
 quadtree-octree 1:500
 unit meshing 1:504–507
 mounted cube drag 3:192–195, 3:196, 3:197–198
 patches 1:477–481
 triangulation 1:599
 SVD *see* singular value decompositions

swept wings 3:346–347
 swirl 3:520–521
 switch functions 3:553
 switch procedures 2:156, 3:351
 symmetric...
 advective-diffusive equations 3:38–40
 coupled FEM/BEM 1:376–389, 1:403–405
 Galerkin boundary element method (SGBEM)
 elasticity 2:727–729
 elastoplasticity 2:727–729, 2:738–740, 2:744
 fracture mechanics 2:733–734
 plastic-elastic analysis 2:727–729, 2:738–740
 Gauss-Seidel implicit time-stepping 3:370–372
 initial value problems 1:680, 1:681
 limited positive (SLIP) scheme 3:351–353, 3:363
 linear elliptic boundary values 1:74–77
 matrix reduced model 2:687–689, 2:691–692
 Navier-Stokes equations 3:416–418
 positive definite matrices 1:555
 reduced models 2:686–689, 2:691–692
 symmetry
 plates 1:302
 Second Moment Closure 3:315
 System-Lax-Friedrichs flux 1:470
 tangents
 matrices 2:219
 moduli tensors 2:14
 stiffness operators 2:337–338, 2:339
 tangential...
 contact stresses 2:204–206
 frictional contact 2:472–473
 sliding 2:199
 velocity 2:198
 tank ships 3:604–605
 Turing's theorem 1:447
 taxonomy 1:543–546, 2:376–377
 Taylor microscale 3:211
 Taylor vortex 3:71–72
 Taylor-Galerkin techniques 3:583
 temperature
 aerocoustics 3:445–446, 3:447
 aircraft cabin air conditioning 3:453
 concrete under fire 2:598
 hypersonic flows 3:433–434
 shakedown 2:307–310
 temporal discretizations 1:331–333
 tensile
 bars 2:345–346, 2:347
 failure 2:438–439
 loadings 2:437–438
 strength 2:532–533
 tension
 bars 2:344–345, 2:346–349, 2:351
 crystal strip necking 2:284
 stiffening 2:529–538
 tensors
 axes 1:538
 Cauchy-Green 2:9–10
 constitutive 2:13–14
 curvature 1:211–212
 damage 2:453–454
 diffusivity 3:139–140
 effective 2:449–451, 2:454, 2:455
 elasticity 2:414–416, 2:449–451, 2:455–456, 2:522–524

fields 1:530, 1:537–538, 2:741
 Finger 3:492
 gradient deformation 2:8–9
 integrity 2:454–456
 Leonard 3:274
 Lighthill turbulence 3:6
 material 2:66–67, 2:449–451, 2:455–456
 metric 1:211–212
 midsurface curvature 1:211–212
 orientation 3:491–492
 products
 adaptive wavelets 1:166
 elements 1:78
 hp-version BEM/FEM 1:393
 spaces 1:121–124
 warped 1:145–146
 window functions 1:282
 Reynolds 3:274, 3:302, 3:303
 rotation 2:77, 2:117–120
 shifter 2:73–74, 2:94–95, 2:98–102
 tangent moduli 2:14
 viscous 3:505
 vorticity correlation 3:210–211
see also strain... stress...
 Terzaghi's theory 2:562, 2:563
 test functions 1:89, 1:119
 testing procedures 2:409–411, 2:417–421
 tetrahedral elements
 error estimates 1:64
 interpolations 2:214
 Maxwell equations 1:728, 1:729–731
 mesh discretization 3:362–363
 refinement 1:101–102, 2:38
 thermal diffusion 1:261–262
 texture maps 1:539
 textures, geometric modeling 1:491–492
 TGS Amira development environment 1:545
 theorem of expended power 2:230–231
 theoretical filters 3:272–274
 thermal...
 diffusion 1:238–240, 1:252–254, 1:257–262
 loadings 2:301
 property upscaling 2:516
 thermo-elastic consolidation 2:588–589
 thermo-hydrodynamics 2:592–599
 thermodynamics 3:506–507
 thermodynamic instabilities 3:503
 thermodynamics
 crystal plasticity 2:272–273
 elasticity balance equations 2:11–13
 equations 2:11–13, 2:230–231, 3:332–334
 first law of 2:12
 forces 2:456
 second law of 2:12 13
 thermomechanical coupling 2:484–485
 thermomechanical loadings 2:293–294, 2:321–326
 thermoplastic strain localization 2:503, 2:504
 thickened wrinkled flame regime 3:518
 thickness
 integration 2:83, 2:98–102
 locking 2:80, 2:97
 plates and shells 1:200–201
 thin domains 1:199–229
 finite element methods 1:219–229
 thin elements 1:228–229

thin plate deflection 1:34–36
 thin sheets 2:495–496
 thin wrinkled flame regime 3:517–518
 thin-walled structures
 delamination-buckling 2:368
 dimensional reduction 2:70–107
 director 2:113–116
 finite element formulation 2:59, 2:104–128
 mathematical modeling 2:61–68
 mechanical foundations 2:63–68
 models 2:59–103
see also plates; shells
 three-dimensional
 aerodynamic shape optimization 3:383–386
 backward extrusion 2:500–502
 blood flow 3:533–540
 computer graphics 1:527–528
 consolidation 2:562
 constitutive models 2:608–609
 contact discretizations 2:212–213
 continuum 2:72–76, 2:83
 design 3:383–386
 discretization 2:198, 2:212–213
 elastic models 2:627–629
 error estimates 1:84–85
 expansions 1:212–213
 finite element models 2:623–624
 flow 3:78–81
 fluid dynamics 3:201–205
 forced homogeneous turbulence 3:225–240
 inviscid flows 3:134–137
 moving boundaries 1:517–521
 Stokes equations 1:266
 thermal diffusion 1:261–262
 three-field methods 2:23, 3:461–464
 three-point bending beams 2:364–365, 2:366–367, 2:369
 thresholding
 adaptive mesh-refining 1:100
 adaptive wavelets 1:172–174, 1:178
 damage 2:482
 visualization algorithms 1:539
 through crack topology 2:388
 through-the-thickness integration 2:94
 through the thickness stretching 2:80
 time
 advancement schemes 3:283
 dataset attributes 1:530
 dependent
 acoustic waves 2:713
 boundary integral equations 1:6, 1:703–719
 compressible Euler equations 3:101, 3:103–104, 3:105
 convection-diffusion equations 1:43–44
 viscoelastic fluid flows 3:481, 3:488–490
 derivatives
 arbitrary Lagrangian-Eulerian 1:418–419
 constitutive equations 2:630–651
 moving volumes 1:418–419
 viscoelastic direct boundary elements 2:752–753
 volume integrals 1:418–419
 discontinuous Galerkin methods 2:173, 2:176–177, 2:184
 discontinuous space-time finite element equations 2:171–173
 discretization
 boundary integral equations 1:715
 dynamic multilevel methods 3:231–240
 fourth-order hyperbolic equations 1:35–36

time (*continued*)
 incompressible viscous flows 3:166–170, 3:175
 nonstationary Navier–Stokes equations 3:166–170
 nonstationary viscous flows 3:175
 Reynolds averaged turbulence 3:242
 shallow water equations 3:242–246
 strong stability preserving Runge–Kutta 3:106–107
 viscoelastic fluid flow 3:494
 domains 1:703–719, 2:713, 2:751–758
 finite element methods 2:176, 2:178, 2:184, 2:186
 harmonic waves 1:707, 1:724–725, 2:697–698, 3:8–9
 increments 1:316
 integration
 discontinuous deformations 1:332–333
 dissipation 2:179, 2:186–187, 3:52–53
 dynamic contact 2:220
 explicit 1:51, 2:183–184, 2:188–189, 3:252–254, 3:256–259
 finite volume schemes 1:464–465
 gravity terms 3:247–248
 incompressible flows 3:67–69
 linear structural dynamics 2:181–184
 multibody contact 1:331–333
 nonlinear aeroelasticity 3:464–465, 3:471–473
 nonlinear structural dynamics 2:184–187
 shallow water equations 3:244–246, 3:252–259
 structural dynamics 2:173–181, 2:188–189
 linear 2:181–184
 nonlinear 2:184–187
 practical considerations 2:187–189
 scale bounds 3:309–310
 scale equations 3:308
 scale separation 3:229–230
 shape-functions 2:754, 2:755–756
 shift invariance 3:275
 splitting 3:67–69
 stepping
 aerodynamics 3:365–379
 boundary integral equations 1:704–705, 1:714–719
 control 1:699, 1:701
 discontinuous deformations 1:332–333, 1:334
 size 2:188–189
 spectral schemes 3:377–379
 stiff initial value problems 1:683–686, 1:687
 turbulent flows 3:225–228
 tissue 2:605–629
 Trefftz structures 1:555, 1:717
 toolkits 1:543, 1:544
 topology
 arbitrary crack growth 2:386–388
 extraction 1:539
 material responses 2:425–426
 shell structure 2:85–86
 validity 1:489–490
 vector fields 1:536–537
 torsion 2:609–613, 2:614, 3:475–476
 total dissipation 3:195–196
 total variation bounded (TVB) Runge–Kutta methods 1:464
 total variational diminishing schemes (TVD)
 computational aerodynamics 3:329
 conservation laws 1:452–453
 critical point accuracy 1:453
 monotone upstream-centered scheme 1:452–453
 multidimensional finite volumes 1:455–456
 one-dimensional finite volumes 1:450–453
 Runge–Kutta methods 1:464, 3:106–107
 shock capturing 3:348

trace operators 1:351
 traction
 acoustic field equations 2:697
 boundary conditions 2:442–443, 2:447, 2:449
 boundary integral equations 1:345–346
 composite laminates 2:444–445
 crystal plasticity 2:286–287
 notched 2:364–365
 plates 1:207
 shakedown 2:293
 three-point bending beams 2:364–365
 vectors 2:10, 2:726
 transfer functions 1:532
 transfer operators 2:489–491
 transfinite mapping 1:421
 transformation methods 2:425
 transforms 2:564
 transient analysis
 buckling 2:160–164
 dynamic processes 1:428
 elastodynamics 2:759, 2:767–768
 Maxwell equations 1:735
 piezoelectricity 2:762, 2:763
 structural dynamics 2:169–189
 transition continuum/discontinuum 1:529–331
 transition elements 2:57–38
 transmission 1:355–356, 1:182, 3:468–471
 transom stern flow 3:591
 transonic
 business jets 3:398–399
 flight 3:390
 flows
 adjoint method 3:382–383, 3:390
 computational aerodynamics 3:326–327
 moving boundaries meshing 1:519–520
 nonlinear aeroelasticity 3:460
 potential 3:537–539, 3:410–411
 small-disturbance equation 3:339–342
 transparency methods 1:301–302
 transpiration condition 3:440–441
 transport
 domain 3:164–166
 element method 3:149–150
 equations 3:310, 3:505–506
 molecular 3:505–506
 neutron 3:91, 3:92–96
 Reynolds 1:419, 3:320
 second moment 3:311–318
 shear stress 3:508–309, 3:311, 3:320
 transverse
 cracking 2:438
 loads 2:438
 normal stiffness 2:126
 normal strains 2:96
 shear
 locking 2:121–123
 strains 2:82, 2:97
 stress 2:97
 trees
 codes 2:143–144
 construction 1:498, 1:500
 objects 1:483–485
 structures 1:189–190
 Trefftz's condition 2:143–144
 Tresca friction 1:433

trial...
 displacements 2:172–173
 and error methods 2:219, 2:642
 functions 1:119
 stress rate 2:256–257
 triangles
 algebraic expansions 1:145–146
 splitting 1:263–264
 to rectangle transformations 3:64–65
 triangular elements
 Hermitian 1:77–78
 Lagrangian 1:77
 macro elements 1:78
 nonconforming 1:105–106
 refinement 1:101, 2:38
 Stokes equations 1:262–266
 thermal diffusion 1:259–260
 triangulation
 curved domains 1:109
 Delaunay-type mesh generation 1:501
 finite element spaces 1:79–80
 geometric modeling 1:478
 mesh refining 1:98, 1:100–103
 numerical integration 1:107–109
 panel clustering 1:599
 triaxial compression 2:555–556
 tribology 2:205
 tridiagonal matrices 1:552
 trigonometric expansion 1:141–142
 trimmed surfaces 1:480, 1:481
 triple decomposition 3:293
 truncation 1:52–53, 1:41, 1:48, 1:611, 3:139–140
 trunk space 1:121–124
 tubes 1:536, 1:540–541, 2:323–326
 tunnels
 autogenous shrinkage 2:514, 2:516–529
 concrete mechanics 2:514, 2:525–529
 concrete under fire 2:598–599
 high-speed train 3:570–571
 linings 2:514, 2:516–529
 turbine flows 3:294
 turbomachinery 3:377–379
 turbulence
 adaptive computation 3:199–201
 aerodynamic flow mechanics 3:418–423, 3:446–447
 closure
 computational fluid dynamics 3:301–322
 constants 3:304–305
 Reynolds averaged 3:318–322
 scalar variable 3:303–311
 second moment transport 3:311–318
 compressible flows 3:271–279, 3:280–281, 3:290–291
 direct numerical simulations 3:2, 3:269–270, 3:279–283, 3:293–296
 dynamic multilevel methods 3:207–264
 flames 3:500–501, 3:514–523
 homogeneous isotropic 3:209–240
 incompressible flows 3:72–74, 3:85–86, 3:271–279, 3:284–290
 industrial aerodynamics 3:418–423, 3:446–447
 kinetic energy 3:446–447
 large eddy simulations 3:2, 3:269, 3:270–274, 3:279–285, 3:293–296
 monoscale schemes 3:419–420
 multiscale method 3:40–53
 Newtonian fluids 3:269–286
 pantograph shrouds 3:5–7
 prediction methods 3:320–322
 Reynolds averaged Numerical Simulations 3:270
 shallow water equations 3:246–247
 ship hydrodynamics 3:593–594
 space-time averages 3:199–201
 spatial behaviors 3:225–228
 spectral/hp element method 3:72–74, 3:85–86
 time behaviors 3:225–228
 time integration 3:247–248
 transport 3:315
 variational multiscale method 3:44–49
 turnkey applications 1:545–546
 TVB *see* total variation bounded
 TVBM property 3:107–108, 3:109
 TVBM stability 3:108
 TVD *see* total variational diminishing
 TVNI schemes 1:451
 two-dimensional
 flows 3:77–78, 3:130–134
 interpolators 2:213–214
 meshes 1:512–513
 moving boundaries 1:517–519
 node-to-segment 2:211–212
 turbulence 3:225–226
 two-field finite element methods 2:21–22
 two-field time-discontinuous Galerkin methods 2:176–177
 two-grid iterations 1:581–584
 two-layer boundary layer model 3:292
 two-layer $k-\epsilon$ model 3:306–307
 two-level decomposition 3:228–232
 two-point boundary problems 1:9–12
 two-point velocity 3:210–211, 3:212
 two-sided Lanczos method 1:573–574
 two-sided preconditioning 1:561
 two-surface yield conditions 2:308–309
 9-schemes 1:21–23, 1:26, 3:489–490, 3:491
 UCAV *see* unmanned combat air vehicles
 UCM *see* upper-convected Maxwell (UCM)
 UCM/RCN stabilization parameters 3:549–550
 UI *see* user interfaces
 unbounded...
 domains 2:702–703
 half-space wave propagation 2:758
 inviscid flows 3:131–137
 uncertainty problem 3:284–285
 uncoagulated test functions 2:709–710
 under-resolved simulations 3:86–87
 uniaxial tension bars 2:344–345, 2:346–349, 2:351
 unidirectional composite laminates 2:445
 uniform...
 bending 2:293
 hierarchical matrices 1:614
 small strain constitutive equations 2:647–649
 test boundary loadings 2:416
 uniqueness 1:30, 2:640–641
 unit meshing 1:504–510
 unit volume 1:507–510
 unity partitions *see* partitions of unity
 universal unfoldings 1:669–670
 unknown ordering 1:563–564
 unmanned combat air vehicles (UCAV) 3:430–431
 unsaturated materials 2:553–554
 unsplit stress-updating 1:429–431
 unstable flames 3:500–501, 3:503–504

- unstable minima 2:645–646
- unsteady advective-diffusive equations 3:36–37
- unsteady Reynolds averaged Navier–Stokes equations 3:321
- unstructured grids 3:63–67
- unstructured meshes 1:457–460, 1:463–464, 3:359–360, 3:396
- updating
 - crystal plasticity 2:274–280
 - dropping 2:279
 - elastic deformation maps 2:275–276
 - iteration 1:657
 - scaling 2:278–279
 - shear 3:555, 3:571, 3:572
 - slip 2:278–280
 - stress 1:428, 1:429–431, 2:274–280
 - thin-walled structures 2:127–128
- upper...
 - bounds 1:97–98
 - convected Maxwell (UCM) 3:481, 3:486, 3:489–490
 - Hessenberg matrices 1:552, 1:561
 - variation bounds 2:414
 - upscaling 2:515–516, 2:521–522, 2:530–531, 2:533–538
- upwinding
 - convection–diffusion equations 1:36–37
 - differentiating 1:36–37, 1:50, 3:330–340, 3:342–343
 - finite difference 1:36–37, 1:50
 - numerical flux 3:92, 3:93
 - shock capturing 3:348–351
 - spectral techniques 3:483
 - triangle schemes 1:463–464
 - user interfaces (UI) 1:475–476
- Uzawa algorithms
 - adaptive wavelets 1:185–186
 - contact mechanics 2:217–218
 - Signorini-type interfaces 1:402–403
- V-cycles 1:585–586, 3:233–234
- v^2-f model 3:317–318
- vacuum-mixed cement pastes 2:519
- validation
 - computational flow mechanics 3:428
 - hierarchical modeling 2:40–42
 - inelastic material constitutive equations 2:639–640
 - nonlinear aeroelasticity 3:474–477
 - van Albeda limiter 1:457
 - Van der Pol's equation 1:685
 - van Leer flux vector splitting 1:470
 - van Leer limiter 1:457
 - vanishing efficiency 3:408–410
 - vanishing viscosity 1:441
 - vapor pressure distribution 2:598
- variable...
 - distribution units 2:667
 - loading domains 2:296–297
 - preconditioning 1:560
 - rank matrices 1:613–614
 - sensitivity methods 2:667–668
- variational
 - alternating Schwarz method 1:619–620
 - equations 3:16–17
 - formulations
 - arbitrary Lagrangian–Eulerian 1:421
 - boundary elements methods 1:347–358
 - boundary integral equations 1:347–358
 - contact mechanics 2:199–200
 - elastic-plastic crystals 2:280–283
- incompressible viscous flows 3:156–157
- Maxwell equations 1:725–727
- microgeometrical manufacturing 2:414
- symmetric Galerkin BEM 2:727–729
- multiscale methods 3:11–18, 3:27–32, 3:38–39, 3:44–49
- principles
 - elasticity 1:242–244
 - mixed finite element methods 1:238
 - partial differential equations 1:293
 - Stokes equations 1:241
 - thermal diffusion 1:239–240, 1:257–262
 - space-time boundary integrals 1:709–711
 - stress updating 2:276–277
 - vascular solid mechanics 2:616–618
 - vaults 2:598–599
 - vdv ordering 1:563
- vecors
 - base 2:8–9, 2:65, 2:91
 - differentiating 2:80, 2:81
 - director 2:118
 - displacement plots 1:534
 - eigenvectors 1:537, 1:552–553
 - fields 1:530, 1:533–537
 - flux splitting 1:469–470, 3:329
 - linear algebraic solvers 1:532
 - load 1:328
 - material heat flux 2:12
 - matrix-vector computations 1:600–604, 1:607, 1:611, 2:15–16, 3:567–568
 - mean value 2:646
 - Ritz 1:553, 1:572
 - rotation 2:118–119
 - Steger–Warming flux splitting 1:469–470
 - traction 2:10, 2:726
 - van Leer flux splitting 1:470
- vein blood flow 3:530
- velocity
 - acoustic field equations 2:697
 - arbitrary Lagrangian–Eulerian 1:416–417, 1:422–423
 - blood flow 3:528–540
 - correction projections 3:67–68
 - homogeneous turbulence 3:239–240
 - Navier–Stokes equations 3:210–213
 - potential 2:697
 - pressure interpolation 3:484
 - shallow water equations 3:241–242, 3:255–259
 - Stokes equations 1:264–266
 - turbulence closure 3:321–322
 - vortex methods 3:141–145
 - VEM see vortex element methods
 - Verfuerth's trick 1:270, 1:272–274
- verification
 - concrete mechanics 2:516, 2:532–533
 - hierarchical modeling 2:40–42
 - inelastic material constitutive equations 2:639
- vertex...
 - centered finite volume methods 1:442, 1:467
 - modes 1:122, 1:123
 - placement 1:541
- vertical pressure amplitude 2:592, 2:593
- vibration
 - eigen-modes 1:200–201, 1:206–207
 - elastodynamics 2:759
 - visbratory response 2:683–692
- VIC see vortex in cell

- Viogt fields 2:409, 2:411–412
- virtual displacements 2:68
- virtual variations 2:199–200
- virtual work
 - degenerating solid elements 2:82–83
 - direct approach 2:79
 - discretization 2:111
 - relations 2:523–524
 - thin-walled structures 2:67–68, 2:74, 2:79, 2:82–83, 2:111
 - three-dimensional continuum 2:74
- viscoelasticity
 - direct boundary elements 2:751–758
 - dynamic analysis 2:751–758
 - fluid flow analysis 3:3
 - Brownian configuration fields 3:491, 3:493–494
 - deformation fields 3:491, 3:492–493
 - flow past cylinders 3:487–488, 3:494–496
 - integral methods 3:490–491, 3:492–493
 - mixed finite element methods 3:481–496
 - numerical methods 3:494
 - steady flows 3:481, 3:482–488
 - stochastic constitutive equations 3:490–491, 3:493–494
 - time dependent flows 3:481, 3:488–490
- viscoplasticity
 - deformations
 - augmented Lagrangian 2:260–263
 - closest-point-projections 2:252–255–32
 - exponential return-mapping 2:248–250
 - integration 2:227–264
 - return-mapping 2:244–250
 - dissipation 2:238–239
- viscosity
 - aerodynamic shape optimization 3:390
 - backing analysis 2:163
 - compressible plasma flows 3:84
 - conservation laws 1:441, 1:466
 - continuum damage models 2:355, 2:369
 - damping 3:305–306
 - incompressible flows 3:155–179
 - modified Lax–Wendroff flux 3:415
 - ship hydrodynamics 3:580, 3:592
 - see also incompressible flows
- viscous...
 - discretization 3:364–365
 - flows
 - aerodynamics 3:330–331
 - ship hydrodynamics 3:585–587
 - vortex methods 3:130, 3:137–139
 - overshoot algorithms 2:278
 - pitching cycles 3:380
 - tensors 3:505
 - viscosity criterion 1:300–303
 - visualization 1:5, 1:525–548
 - algorithms 1:531–541
 - data forms 1:528–531
 - graphics 1:527–528
 - interfacing 1:546–548
 - large data methods 1:542–543
 - taxonomy 1:543–546
 - volume rendering 1:541–542
- VOF see volume of fluid
- volume
 - domains 1:499
 - of fluid (VOF) 3:583
 - fractions 2:418–421, 2:425–426, 2:514–515, 2:520–521
- integrals 1:418–419
- potentials 1:704
- rendering 1:541–542
- residuals 1:87–88
- volumetric...
 - behavior, geomechanics 2:557, 2:558
 - flow rates 3:533, 3:534, 3:539–540
 - locking 1:298–300
- von Mises effective stress distribution 2:475–476
- von Mises yield criterion 2:234–235
- Voronoi cells 2:425, 2:440
- vortex...
 - blobs 3:131, 3:132
 - bursting 3:468–469
 - in cell (VIC) method 3:130, 3:144–145
 - drag 3:443
 - element methods (VEM) 3:131–134, 3:145–149
 - filaments 3:134–135
 - methods 3:2, 3:129–152
 - bluff-body flows 3:131–137
 - convection–diffusion equations 3:149
 - efficient velocity evaluations 3:141–145
 - flows 3:145–149
 - incompressible flows 3:131–137
 - inviscid flows 3:131–137
 - particle redistribution 3:140–141
 - truncation errors 3:139–140
 - viscous flows 3:137–139
 - particles 3:134, 3:135–137, 3:149–150
 - sheets 3:131, 3:133–134, 3:151
- vorticity
 - correlation tensors 3:210–211
 - cut 3:131, 3:152–153
 - error indicators 3:174–175
 - military aircraft 3:429–431
 - visualization algorithms 1:536
 - vortex representation 1:476–477, 1:484–485, 1:532, 1:533
 - VTK toolkit 1:544
- W-cycles 1:585–586, 3:233
- wake 3:72–74, 3:598–599
- wall...
 - boundary conditions 3:291–292, 3:296, 3:361
 - bounded flows 3:426
 - echo 3:316–317, 3:320–321
 - functions 3:305–306, 3:307–308
 - impedance conditions 2:689
 - normal displacements 2:685, 2:687, 2:690, 2:691
 - shear stress 3:537–538
 - temperatures 3:453
- warped tensor products 1:145–146
- water
 - falling solids 3:604–607
 - flow 2:584–586
 - inside hulls 3:581
 - mills 3:603
 - pore 2:561–562
 - shallow water equations 3:240–259
- wave...
 - contact 3:604–605
 - drag 3:443
 - equation
 - derivation 1:28–29
 - difference approximation 1:30–34
 - discontinuous Galerkin methods 3:94

- wave... (continued)
- domain of dependence 1:29–30
 - finite difference methods 1:28–34
 - Helmholtz equation 2:701–702
 - Maxwell equations 1:724–725
 - space-time boundary integrals 1:706–707, 1:711–713
 - time-harmonic waves 2:697–698
 - guides 1:735
 - patterns 3:580, 3:597–598
 - propagation 2:757–758, 3:531–533
 - resistance 3:579, 3:580
 - ship contacts 3:601, 3:602–604
 - transmission 3:6–7
- waveform advection 3:62
- wavelets
- adaptive techniques 1:157–195
 - elasticity 2:731
 - Galerkin schemes 1:178–180
 - Haar 1:160–162, 1:481
 - linear operators 1:168–169
- wavenumbers
- accelerated multifrequency methods 2:710–711
 - eddy viscosity 3:47–49
 - element Green's function 3:26
 - Helmholtz equation 2:699–701, 3:26
 - time-harmonic waves 2:697–698
- weak form
- boundary...
 - elements methods 1:347–358
 - integral equations 1:347–358
 - values 1:446–447, 2:16–17
 - conservation laws 1:440–441
 - contact mechanics 2:200–202
 - continuity equation 1:726
 - Euler equation solutions 3:413
 - Godunov finite volume discretizations 1:444
 - linear elliptic boundary values 1:74–75
 - matrix compressions 1:177–178
 - partial differential equations 1:293–294
- weakly singular kernel integral equations 1:594–595
- wear 2:206
- weighted...
- averages 1:30–31
 - essentially nonoscillatory (WENO) schemes 1:455, 3:355–356
 - least squares 2:643
- residuals 2:176, 2:177, 2:178
 - Sobolev spaces 1:63
- weighting functions
- exterior acoustics problems 2:708–709
 - moving least squares 1:290–291
 - smooth particle hydrodynamics 1:281–283, 1:284
 - structural dynamics 2:172–173
 - Weissenburg numbers 3:481, 3:485–488, 3:489–490, 3:494–496
- well-posedness
- adaptive wavelets 1:168–169, 1:184–185
 - discontinuous Galerkin methods 3:97
 - finite difference methods 1:7–9
 - wavelets and linear operators 1:168–169
- WENO *see* weighted essentially nonoscillatory
- Whitney space 1:730
- Wigley hulls 3:594–597
- wind loads 2:533–534, 2:538
- wind tunnels 3:293, 3:429–430
- window functions 1:281–283, 1:284
- winged-edge data structures 1:481, 1:483
- wings 3:78–81, 3:392–397
- wire-basket-based Schur-complement preconditioning 1:636
- World War II 3:408–409
- wrapping lines with tubes 1:536, 1:540–541
- X-38 crew rescue vehicles 3:434–435
- XFEM *see* extended finite element method
- yield
- conditions 2:227–228, 2:233–235, 2:308–309
 - functions 2:233, 2:310, 2:465
 - line patterns 1:422
 - shakedown 2:308–309
 - strains 2:481
 - surfaces 2:227–228, 2:234–235, 2:481
 - Young's modulus 2:559–560
- Zarka's method 2:319–320
- zero-dimensional tools 3:513–514
- zero-frequency mode 2:686, 2:691
- zeroth-order approximations 2:449–451
- Zienkiewicz elements 1:78

